REDACTED – FOR PUBLIC INSPECTION

September 25, 2019

Rebekah Goodheart
Tel +1 202 639 5355
RGoodheart@jenner.com

VIA ECFS

Ms. Marlene H. Dortch
Secretary
Federal Communications Commission
445 12th Street, SW
Washington, DC 20554

Re: *Telecommunications Relay Services and Speech-to-Speech Services for Individuals with Hearing and Speech Disabilities*, CG Docket No. 03-123

Dear Ms. Dortch:

CaptionCall, LLC herein submits a REDACTED version of its comments on the three pending applications for certification to provide Internet Protocol Captioned Telephone Service.

CaptionCall is submitting a Highly Confidential version pursuant to the *Third Protective Order* adopted for the above-captioned docket.[1] CaptionCall has received written approval from staff to designate for Highly Confidential treatment the marked portions of the comments, which include information designated as Confidential or Highly Confidential by the filing parties in the pending applications.[2]

Pursuant to the *Third Protective Order*, CaptionCall is submitting the Highly Confidential version for the Secretary and two copies for Eliot Greenwald.

Please contact me if you have any questions or require any additional information.

Sincerely,
/s/ Rebekah. P. Goodheart
Rebekah P. Goodheart
Counsel for CaptionCall, LLC

[1] *See In re Telecommunications Relay Services and Speech-to-Speech Services for Individuals with Hearing and Speech Disabilities*, Order and Third Protective Order, CG Docket Nos. 03-123, 10-51, 13-24, DA 18-751 (2018) ("*Third Protective Order*").

[2] *Third Protective Order* Appendix B.

Enclosures

REDACTED – FOR PUBLIC INSPECTION

Before the
FEDERAL COMMUNICATIONS COMMISSION
Washington, D.C. 20554

In the Matter of	)	
Telecommunications Relay Services and Speech-to-Speech Services for Individuals with Hearing and Speech Disabilities	)	CG Docket No. 03-123

COMMENTS OF CAPTIONCALL, LLC ON THE APPLICATIONS OF MACHINEGENIUS INC., VTCSECURE, LLC, AND CLARITY PRODUCTS, LLC FOR CERTIFICATION TO PROVIDE AUTOMATIC SPEECH RECOGNITION BASED INTERNET PROTOCOL CAPTIONED TELEPHONE SERVICES

Rebekah P. Goodheart
Elliot S. Tarloff
James T. Dawson
JENNER & BLOCK LLP
1099 New York Avenue, NW
Suite 900
Washington, DC 20001
(202) 639-6000

Counsel for CaptionCall, LLC

September 25, 2019

TABLE OF CONTENTS

REDACTED – FOR PUBLIC INSPECTION

Before the
FEDERAL COMMUNICATIONS COMMISSION
Washington, D.C. 20554

In the Matter of	)	
Telecommunications Relay Services and Speech-to-Speech Services for Individuals with Hearing and Speech Disabilities	)	CG Docket No. 03-123

COMMENTS OF CAPTIONCALL, LLC ON THE APPLICATIONS OF MACHINEGENIUS INC., VTCSECURE, LLC, AND CLARITY PRODUCTS, LLC FOR CERTIFICATION TO PROVIDE AUTOMATIC SPEECH RECOGNITION BASED INTERNET PROTOCOL CAPTIONED TELEPHONE SERVICES

CaptionCall, LLC ("CaptionCall") submits these comments on the applications to provide Internet Protocol Captioned Telephone Service ("IP CTS") using automatic speech recognition ("ASR") technology filed by MachineGenius Inc. ("MachineGenius"), VTCSecure, LLC ("VTC"), and Clarity Products, LLC ("Clarity").[1]

INTRODUCTION AND SUMMARY

CaptionCall is committed to innovation and believes that ASR holds tremendous promise. The three pending ASR applications, however, fail to meet the burden of showing that their services satisfy the standards for certified telecommunications relay services ("TRS") under

[1] *See Comment Sought on Application of Clarity Products, LLC for Certification to Provide Internet Protocol Captioned Telephone Service*, Public Notice, CG Docket No. 03-123, DA 19-820 (rel. Aug. 26, 2019); *Comment Sought on Application of MachineGenius, Inc. for Certification to Provide Internet Protocol Captioned Telephone Service*, Public Notice, CG Docket No. 03-123, DA 19-819 (rel. Aug. 26, 2019); *Comment Sought on Application of VTCSecure, LLC, for Certification to Provide Internet Protocol Captioned Telephone Service*, Public Notice, CG Docket No. 03-123, DA 19-818 (rel. Aug. 26, 2019).

the Americans with Disabilities Act (the "ADA")[2] and the Commission's mandatory minimum standards.[3] In addition, the Commission lacks a framework for evaluating ASR-based IP CTS services, including standards that reflect the differences between existing certified TRS services and ASR-based services. The Commission must adopt such a framework before the agency is able to accurately evaluate whether ASR-based services comply with the ADA.

In the Commission's declaratory ruling finding that ASR-based IP CTS can be TRS (the "*Declaratory Ruling*"), the Commission described the applicability of the mandatory minimum standards to ASR-based IP CTS and indicated that "applicants should support all claims regarding their use of ASR and its efficacy through documentary and other evidence."[4] In response, several parties on the record, including the Commission's Disability Advisory Committee, the Consumer Groups, the Clear2Connect Coalition, Sprint, and Hamilton, urged the Commission to adopt a qualitative and quantitative framework for the evaluation of applications to provide ASR-based IP CTS, to ensure that individuals with hearing loss do not receive services that are not capable of enabling functionally equivalent communications by telephone.[5]

---

[2] *See* 47 U.S.C. § 225 (requiring that TRS enable "functionally equivalent" communications by telephone).

[3] *See* 47 C.F.R. § 64.604; *id.* § 64.606(a)(2).

[4] *See In re Misuse of Internet Protocol (IP) Captioned Telephone Service*, Report and Order, Declaratory Ruling, Further Notice of Proposed Rulemaking, and Notice of Inquiry, 33 FCC Rcd 5800, 5832-35 ¶¶ 60, 63 (2018) ("*Declaratory Ruling*").

[5] *See, e.g.*, Recommendation of the FCC Disability Advisory Committee, Relay and Equipment Distribution Subcommittee: Internet Protocol Captioned Telephone Relay Service Metrics ¶ 6 (adopted Oct. 3, 2018), https://docs.fcc.gov/public/attachments/DOC-354522A1.pdf; Letter from Blake E. Reid, Counsel to Telecommunications for the Deaf and Hard of Hearing, Inc., to Marlene H. Dortch, Secretary, FCC, CG Docket Nos. 03-123 and 13-24, at 3-7 (July 26, 2018) ("*Consumer Groups Framework*"); Letter from Clear2Connect Coalition to Marlene H. Dortch, Secretary, FCC, CG Docket Nos. 13-24, 03-123 (May 14, 2019) ("*C2C Ex Parte*"); Comments of Hamilton Relay, Inc., CG Docket Nos. 13-24, 03-123, at 13 (Sept. 7, 2018) ("*Hamilton PFR Comments*"); Sprint Petition for Clarification or, in the Alternative, Reconsideration, CG Docket Nos. 13-24, 03-123, at 2-3 (July 9, 2018) ("*Sprint PFR*"); *accord* Reply Comments of CaptionCall, CG Docket Nos. 13-24, 03-123, at 2, 6 (Nov. 15, 2018) (discussing importance of establishing "consistent, objective, and technology neutral" service

With the guidance from the attached report from Dr. Stern—a Professor in the Department of Electrical and Computer Engineering, the Language Technologies Institute, and the Department of Computer Science at Carnegie Mellon University[6]—CaptionCall provides recommendations for such a framework based on the *Declaratory Ruling*, which can be adjusted after the Commission completes the process initiated by its notice of inquiry to adopt IP CTS performance standards (the "NOI").[7]

Even if the Commission does not adopt such a framework, however, each of the currently pending applications falls short of demonstrating that its service satisfies the federal standards, and, accordingly, each should be denied without prejudice.

I. The Commission Should Adopt a Framework for Certifying ASR-Based IP CTS Providers Built Around the Mandatory Minimum Standards as Explained in the *Declaratory Ruling* and the Standard of the ADA.

In the *Declaratory Ruling*, the Commission outlined the certification process for ASR-based IP CTS providers. There, the Commission explained that applications must establish that their services meet the Commission's mandatory minimum standards and, among other things, (1) show how their service is capable of handling "all types of calls," (2) show how the provider will "ensure that conversations are kept confidential," and (3) "demonstrate that their services

quality metrics, testing methodologies, and standards that should be applied to all prospective IP CTS providers). CaptionCall notes that the *Sprint PFR* of the *Declaratory Ruling* remains pending.

[6] Dr. Stern's Report is attached to these Comments as Appendix A.

[7] *See Declaratory Ruling*, 33 FCC Rcd at 5870-73 ¶¶ 164-170. As the Commission is aware, existing IP CTS providers are making progress toward the development of uniform service quality metrics, testing methodologies, and service quality standards that could be applied to all IP CTS providers in a technology neutral manner. *See* Letter from Dixie Ziegler, Cristina Duarte, Michael Strecker, Bruce Peterson, Scott Freiermuth, and Kevin Colwell to Marlene H. Dortch, Secretary, FCC, CG Docket Nos. 13-24, 03-123 (Sept. 20, 2019) ("*Industry Update*"); Letter from Bruce Peterson, Dixie Ziegler, Cristina Duarte, Michael Strecker, and Scott Freiermuth to Marlene H. Dortch, Secretary, FCC, CG Docket Nos. 13-24, 03-123, Attach. (Aug. 21, 2018).

support 911 emergency calling."[8] The Commission also directed applicants to "support all claims regarding . . . [ASR's] efficacy through documentary and other evidence," including "quantitative test results demonstrating that the applicant's service will afford a level of quality that is at least comparable to currently available CA-assisted IP CTS with respect to captioning transcription, delays, accuracy, speed, and readability."[9]

The Commission directed the Bureau not to approve any application unless the provider shows how it can meet these standards.[10] Given the current state of ASR, however, it is unclear if ASR-only providers are able to generate captions that are "at least comparable to currently available CA-assisted IP CTS,"[11] especially with respect to difficult calls.[12] As Dr. Stern explains, ASR-only services are more likely to struggle during emergency calls;[13] calls with difficult speakers (*e.g.*, soft-spoken speakers, speakers with unusually high- or low-pitched

---

[8] *Declaratory Ruling*, 33 FCC Rcd at 5832-33 ¶ 60.

[9] *Id.* at 5834 ¶ 63.

[10] *See id.* at 5834 ¶ 63 ("We also make clear that the Bureau will not approve any application to provide IP CTS using ASR that does not demonstrate that the applicant will meet the Commission's mandatory minimum standards for functional equivalence . . . . [A]ny certification application relying on . . . deficient technology will not be approved.").

[11] *Id.*

[12] If a provider has sufficient CAs to handle all *types of calls* and, indeed, all *calls* (as might be necessary when the ASR engine does not perform adequately across multiple calls or otherwise fails), these concerns would be minimized. But the Commission should exercise caution with new entrants with plans to offer "hybrid" services. If a certified "hybrid" provider has only a few CAs, it will not be able to deliver functionally equivalent communications if its ASR is unable to perform, or performing inadequately, during multiple simultaneous calls. Thus, for example, while VTC plans to "utilize[] a [h]ybrid approach that allows for traditional use of [CAs]," VTCSecure, LLC Request for Waiver, CG Docket No. 03-123, at 2 (Sept. 13, 2019) ("VTCSecure Waiver"), **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]** The Commission should require any hybrid ASR applicant to submit information for its CA-based operations and show that it has plans to hire and train enough CAs given the nature of its proposed hybrid service.

[13] *See* Stern Report at 10, 14.

voices, minority speakers, speakers with accents, and speakers with speech impairments);[14] calls with difficult speech content (*e.g.*, speech content that is highly specialized or personalized);[15] calls with difficult background conditions (*e.g.*, noise, music, multiple voices, or reverberation);[16] or calls when captioning is otherwise inaccurate for unidentifiable reasons. Moreover, the Commission should ensure that applicants are capable of satisfying the requirements *at scale* (demonstrated by submission of load data) and *over time* (demonstrated by submission of data from over a period of time). An ASR-only service that performs well during a single demonstration under perfect conditions may not perform well when handling thousands of calls, and hundreds of thousands of minutes, week after week.[17]

Below, and based on the attached report of Dr. Richard Stern, CaptionCall proposes the following framework for evaluating the pending applications.[18]

---

[14] *See id.* at 11-15.

[15] *See id.* at 16-17.

[16] *See id.* at 15-16.

[17] *See* 47 C.F.R. § 64.604(b)(4) (requiring providers to offer service 24 hours per day); *Declaratory Ruling*, 33 FCC Rcd at 5832-33 ¶ 60 (applying requirement to prospective ASR-only providers).

[18] As previously explained, the Commission also lacks an adequate record to adopt an ASR-only compensation rate. There is *no* support on the record for the Commission's proposed rate of $0.49 per minute, which, as CaptionCall and others have explained, is the result of arbitrary and capricious exclusions of certain costs by Rolka Loube, as well as the arbitrary and capricious categorization of cost categories as either fixed or variable. *See* Reply Comments of CaptionCall, LLC, CG Docket Nos. 13-24, 03-123, at 23-24 (Oct. 16, 2018) (describing "near unanimity . . . that the Fund Administrator's calculation [of an ASR-only rate] suffers from evidentiary and methodological problems that improperly depressed the recommended rate" and that would "not provide the proper incentives for at-scale providers to develop and use ASR"). The Commission should not certify any ASR-only provider until it has determined the appropriate compensation methodology for such providers.

A. Documentary and Other Evidence

The Commission's rules require any TRS provider to submit "[a] detailed description" of how it will meet mandatory minimum standards "including documentary and other evidence."[19] In the *Declaratory Ruling*, the Commission made clear that ASR-based IP CTS applicants should provide such evidentiary support for "all claims regarding the[] use of ASR and its efficacy."[20] For example, the Commission explained that an applicant could "include trials and quantitative test results" demonstrating that the service matches the quality of current CA-based services.[21] To satisfy the requirement of providing sufficient documentary evidence, an application should, at a minimum, (1) identify and describe the testing methodology,[22] evaluation protocols, and results on which the provider is relying; and (2) all such testing should have been performed, or reviewed and certified, by a third party.

First, the Commission should require providers to explain how the ASR service was tested and evaluated (the methodologies and evaluation protocols), and how it performed (the results) during the testing.[23] Absent such information, the Commission will not be able to assess whether the service was tested using an appropriate methodology and is therefore capable of

---

[19] 47 C.F.R. § 64.606(a)(2)(ii).

[20] *Declaratory Ruling*, 33 FCC Rcd at 5834 ¶ 63.

[21] *Id.* at 5834 ¶ 63.

[22] This requirement would include providing descriptions of the speakers and types of calls that were used for testing purposes, but would not include all of the raw results generated from the tests. Dr. Stern refers to this as testing "datasets."

[23] In reviewing any application to be a hybrid provider, this process would encompass providing results for both the provider's CA-based performance and its ASR-based performance to show that its hybrid system as a whole enables functionally equivalent communications—as well as the testing methodologies and evaluation protocols that were used in its testing.

handling all types of calls, nor will the Commission know how ASR-based services compare to existing CA-based services in terms of accuracy, readability, etc.

Second, there is support in the record from a variety of stakeholders that the testing to evaluate IP CTS providers, including ASR-based providers, must be objective and generalizable.[24] As Dr. Stern describes, deferring certification until quality metrics have been adopted would drive innovation faster over the long run, while also protecting consumers from providers who over-promise and under-deliver. At a minimum, the Commission should require that all testing submitted to support an application for certification to provide ASR-only IP CTS be conducted, or reviewed and certified, by a third party. This will help guard against providers' performing tests under conditions to achieve predetermined outcomes and/or in scenarios that do not reflect real-world calling scenarios.

Third, the Commission should not rely on MITRE's test results as sufficient evidence to find that any pending ASR applications satisfy the Commission's standards. Even if MITRE's test results established that ASR-based services *in general* are capable of delivering functional equivalence—which they do not—the Commission would lack a basis for concluding that *a particular applicant* will satisfy the Commission's mandatory minimum standards or that *its specific* ASR product will deliver functional equivalence. Indeed, even if the Commission could confirm that a particular applicant's underlying ASR engine had been part of the MITRE tests,

[24] *See, e.g.*, *Industry Update* at 3; *Consumer Groups Framework* at 4-5; *C2C Ex Parte* at 2-3. For example, if a provider tests an ASR system by sending recorded speech directly to the ASR's application programming interface, it might achieve an idealized result. The burden should be on the applicant to demonstrate that accuracy was tested in conditions that resemble its production environment to allow the Commission to draw conclusions from the testing results.

the Commission would still lack evidence that the applicant's processes for using that engine to provide service are adequate.[25]

Fourth, it is not sufficient for applicants to agree to undergo future testing. Clarity, for example, has responded to concerns that it has not conducted sufficient testing by stating that it "is fully ready, and, indeed, eager to prove that its [ASR-based IP CTS offering] will surpass all existing IP CTS providers in speed, accuracy, and ease of use."[26] But the Commission should not certify a provider based on its future promise; instead, it should require all applicants to submit evidence to support their claims, which, as discussed below, none of these applicants has done.[27]

B. All Types of Calls

The mandatory minimum standards require that certified TRS providers be capable of handling "any type of call normally provided by telecommunications carriers."[28] It is critical that the Commission ensure this standard is met, because, as Dr. Stern explains, there is evidence

---

[25] Even if the Commission could impute MITRE's results to any individual provider—which it cannot, as discussed above—MITRE's results do not establish that ASR-only services will satisfy the mandatory minimum standards as clarified in the *Declaratory Ruling*. *See* Stern Report at 22-24; *accord* Letter from David. A. O'Connor, Counsel for Hamilton Relay, Inc., to Marlene H. Dortch, Secretary, FCC, CG Docket Nos. 13-24, 03-123, at 2-3 & n.6 (May 24, 2018) ("*Hamilton 5-24 Ex Parte*"); *C2C Ex Parte* at 11; Letter from John T. Nakahata, counsel to CaptionCall, LLC, to David Schmidt, TRS Fund Program Coordinator, Office of Managing Director, FCC, CG Docket Nos. 03-123, 13-24 Attach. 1 (Dec. 21, 2017).

[26] Letter from Seymour James van den Bergh, CEO, Clarity Products, LLC, to Marlene H. Dortch, Secretary, FCC, CG Docket Nos. 03-213 and 13-24, at 3 (May 30, 2019) ("*Clarity May 30 Letter*").

[27] Clarity argues that "testing should not delay certification." *Id.* at 1; *see id.* at 6 (suggesting that Clarity "welcomes the opportunity to conduct additional testing" but that "such efforts should not delay Commission certification"). Instead of providing the types of quantitative test results contemplated by the *Declaratory Order*, however, Clarity instead says that it "believe[s]" that it is "likely that today's ASR technologies will . . . perform better than CA-based IP CTS" and claims that it "would welcome evidence to the contrary." *Id.* at 5. This argument shifts the burden of proof established in the *Declaratory Ruling*. Moreover, no one can submit evidence regarding a specific ASR-based service unless its provider includes the testing methodologies, and protocols that are being relied upon in the application to demonstrate compliance with the mandatory minimum standards and the ADA.

[28] 47 C.F.R. § 64.604(a)(3).

that ASR-only services will not be able to handle all calls equally.[29] And without the ability to fall back on trained and professional CAs to provide and correct captions during such calls, a certified ASR-only IP CTS provider may generate inaccurate captions that do not enable functionally equivalent communications.[30]

The Commission should ensure that each pending application satisfies the "all types of calls" requirement. The most objective way to confirm that each service can handle all types of calls would be to test them all using the same testing methodology.[31] But at a minimum, every applicant should explain what testing methodology, evaluation protocol, and results it is relying upon to show that the service is effective for all types of calls, regardless of the identity of the speakers, the subject matter of the call, or the background conditions.[32]

These requirements would provide the Commission with data to evaluate whether any ASR-based services feature algorithmic bias. The next generation of artificial-intelligence technologies must be trained with representative data, or else they will perform less well for certain under-represented and minority groups.[33] The Commission has the authority it needs under the ADA to prevent this from occurring in this proceeding, and it should do so: Many of

---

[29] *See supra* notes 13-16 and accompanying text. An individual without hearing loss might experience more difficulty for some calls than others. But those incremental differences in call quality are also reflected in CA-based IP CTS. As explained in the Stern Report, however, ASR-only services are likely to provide *systematically and significantly worse* quality over and above the degraded quality that might be expected in either a normal telephone call or a CA-based IP CTS call.

[30] *See supra* note 12.

[31] *See Industry Update* at 3 ("Testing must be performed and reported over a wide variety of call types. . . . Failure to test, score, and report on a variety of call types, involving a variety of callers and call conditions, risks harm to specific populations, including, but not limited to, children, older adults, those with speech disabilities, and other individuals with less common accents.").

[32] For the same reasons discussed above, all such testing should have been performed, or reviewed and certified, by a third party.

[33] *See, e.g., C2C Ex Parte* at 4 & n.20, 10 & n.42.

the types of speakers for whom ASR-only services will perform least well are groups that receive status-based legal protections because they are historically vulnerable and/or subject to discrimination; it would be inconsistent with the spirit and requirements of the ADA, a federal civil rights law, to introduce such biased services into a market that is currently serving its users well.

C. 911 and Public Safety Calls

The Commission clarified that an applicant to provide ASR-only IP CTS must "*demonstrate* that [its] service support[s] 911 emergency calling and meets applicable emergency call handling requirements."[34] The Commission did so noting its prior history of certifying IP CTS providers who failed to operate the service in compliance with emergency calling requirements—including providers who disclaimed any responsibility for 911 call handling.[35] The Commission likewise requested comment on whether there are "unique challenges with respect to relaying calls to 911 associated with any of the methods used to generate IP CTS captions" including ASR-based systems.[36] The Commission did so due to concerns about the ability of ASR-only providers that provide or contract for public switched network transmission to route 911 calls to appropriate PSAPs without dropping them. In addition, the Commission must account for evidence that ASR-only services will face difficulties

[34] *Declaratory Ruling*, 33 FCC Rcd at 5833 ¶ 60 (emphasis added); *see also* 47 C.F.R. § 64.605.

[35] *Declaratory Ruling*, 33 FCC Rcd at 5833 ¶ 60 & n.208; *see also In re Telecommunications Relay Services and Speech-to-Speech Services for Individuals with Hearing and Speech Disabilities*, Order, 29 FCC Rcd 13,716 (CGB 2014); *In re Misuse of Internet Protocol (IP) Captioned Telephone Service*, Order, 30 FCC Rcd 2934 (CGB 2015) ("*InnoCaption 911 Order*").

[36] *Declaratory Ruling*, 33 FCC Rcd at 5867 ¶ 153.

in generating accurate captions during highly stressful 911 and other emergency calls, as Dr. Stern explains.[37]

To ensure that emergency calls are handled in a way that is functionally equivalent, the Commission should require each applicant to allow Commission staff to test its ability to handle and route 911 calls as part of the certification process.[38] The Commission previously has engaged in such testing *after having certified* a provider that had certified as to its ability to handle 911 calls.[39] But given the untested nature of this technology, such an ex post approach is inadequate here. Likewise, because of the unique risks associated with dropped or mishandled 911 calls, applicants should be required to identify and describe their failover plans for 911 calls.[40] Their failover plans, too, should be subject to testing prior to certification.

D. Confidentiality and Privacy

In the *Declaratory Ruling*, the Commission concluded that ASR will enhance user privacy, because captions can be generated without the involvement of a CA.[41] But the record

---

[37] *See* Stern Report at 10, 14; *accord Hamilton 5-24 Ex Parte* at 2. Concerns about public safety extend beyond 911 call handling and routing. Thus, the Commission should apply CaptionCall's recommendations in this section—regarding testing and failover plans—not just to 911 calls, but also to other emergency and public safety calls. Indeed, recent research performing a "systematic comparison of selected ASRs" for speech between clinicians and patients in clinical scenarios "likely to be seen in an ambulatory primary care practice" found word error rates "ranging from 65% to 34%, all falling short of the rates achieved for other conversational speech." Jodi Kodish-Wachs et al., *A Systematic Comparison of Contemporary Automatic Speech Recognition Engines for Conversational Clinical Speech* (2018), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371385/. The authors concluded that the "modest level of performance suggests that we need to focus on improving ASR engine performance before we can adopt these technologies for conversational speech for a broad range of clinical use cases." *Id.*

[38] *See, e.g.*, *C2C Ex Parte* at 9; *Consumer Groups Framework* at 5.

[39] *InnoCaption 911 Order*, 30 FCC Rcd at 2937-38 ¶¶ 5-6.

[40] Here too, having a sufficient number of trained CAs would be a minimum safeguard to ensure that all such emergency calls are captioned at current levels of quality.

[41] *Declaratory Ruling*, 33 FCC Rcd at 5828 ¶ 50.

reveals that this conclusion is overly simplistic in two respects.[42] The Commission thus needs to ensure that ASR-based providers also protect user confidentiality and privacy.

First, the Commission's rules protect the privacy of traditional, CA-based IP CTS users. Indeed, the Commission has described that "[r]elay services are unique in that, in the present technological environment, they utilize human CAs who see and hear private conversations while acting as transparent conduits relaying conversations without censorship or monitoring functions."[43] The Commission's rules require notification that CAs are present on phone calls,[44] and the Commission's rules prohibit certified providers from "disclosing the content of any relayed conversation" and "keeping records of the content of any conversation."[45]

Second, the absence of a CA does not necessarily enhance user privacy. The Commission has noted that "in an effort to improve ASR accuracy, certain companies are engaged in research that applies network algorithms to the content of users' speech, which is sometimes captured in the Internet 'cloud.'"[46] Subsequent events confirm that these companies are doing much more than that. It has been widely reported that many of the companies that could or likely would provide wholesale ASR to certified IP CTS providers have recorded audio

---

[42] *See C2C Ex Parte* at 3-4.

[43] *In re Telecommunications Services for Individuals with Hearing and Speech Disabilities, and the Americans with Disabilities Act of 1990*, Report and Order and Request for Comment, 6 FCC Rcd 4657, 4659 ¶ 13 (1991).

[44] 47 C.F.R. § 64.604(c)(11)(v).

[45] *Id.* § 64.604(a)(2).

[46] *Declaratory Ruling*, 33 FCC Rcd at 5834 ¶ 63; *see also Hamilton PFR Comments* at 11.

content to train speech recognition engines—including by allowing their employees or contractors to listen to the audio to evaluate transcriptions.[47]

The Commission "ask[ed] applicants for ASR certification to provide information about measures they will take to ensure the confidentiality of call content transcribed through an automated speech engine to ensure compliance" with its confidentiality rules.[48] The Commission must confirm, however, that the "measures" taken satisfy the mandatory minimum standards—not just that those measures otherwise protect user privacy. Additionally, to enhance transparency and ensure that consumers can make informed decisions, the Commission should also require applicants who intend to rely on a third-party ASR provider to identify that third party publicly.

II. No Pending Application Satisfies the Commission's Requirements.

Three entities have filed applications with the Commission to provide ASR-based IP CTS: MachineGenius (filed October 2017), VTC (filed May 2017), and Clarity (filed April 2019). The Commission has explained that ASR "applicants should support all claims regarding their use of ASR and its efficacy through documentary and other evidence" and this requirement could be satisfied by "trials and quantitative test results demonstrating that the applicant's service will afford a level of quality that is at least comparable to currently available CA-assisted

---

[47] *See C2C Ex Parte* at 4; *see also, e.g.*, Maria Armental, *Apple Tightens Privacy Rules on Siri Recordings After Backlash*, Wall St. J. (Aug. 28, 2019), https://www.wsj.com/articles/apple-tightens-privacy-rules-on-listening-to-siri-recordings-11567013482; Sarah E. Needleman & Parmy Olson, *Google Contractors Listen to Recordings of People Using Virtual Assistant*, Wall St. J. (July 11, 2019), https://www.wsj.com/articles/google-contractors-listen-to-recordings-of-consumers-addressing-virtual-assistant-11562865883; Geoffrey A. Fowler, *Alexa Has Been Eavesdropping on You This Whole Time*, Consumer Tech Perspectives, Wash. Post (May 6, 2019), https://www.washingtonpost.com/technology/2019/05/06/alexa-has-been-eavesdropping-you-this-whole-time/.

[48] *Declaratory Ruling*, 33 FCC Rcd at 5834 ¶ 63.

IP CTS with respect to captioning transcription delays, accuracy, speed, and readability."[49] As explained below, none of the pending applications meets this standard or provides sufficient evidence that the provider's offering will comply with the Commission's rules. For that reason, each of the pending applications should be denied, without prejudice to the applicant reapplying in the future if and when it can demonstrate that its service will comply with the governing standards.

A. MachineGenius, Inc.

MachineGenius states that it is a "recently-formed technology start-up corporation" based in Massachusetts.[50] MachineGenius's IP CTS offering is "Olelo," which is an over-the-top app that can be installed on mobile devices such as smartphones and tablets.[51] Olelo will provide transcription using ASR only, and MachineGenius will not employ any CAs to assist with call transcription.[52] MachineGenius's application—which provides no details at all concerning the testing of Olelo—falls well short of the Commission's standards.[53]

According to the application, Olelo **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted]

---

[49] *Id.*

[50] MachineGenius, Inc. Internet-Based TRS Certification Application, CG Docket 03-123 at 4 (Oct. 13, 2017) ("MachineGenius App.").

[51] *Id.* at 5.

[52] *Id.* at 7. **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]** *Id.* at 10.

[53] *See Declaratory Ruling*, 33 FCC Rcd at 5832-35 ¶¶ 60, 63.

[54] MachineGenius App. at 7-8.

[redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**[55] That said, MachineGenius's application provides no details on how **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**[56]

Moreover, MachineGenius's application does not include "documentary" or "other evidence" like "trials and quantitative test results" confirming the ability to match existing service quality levels, as is contemplated by the *Declaratory Ruling*. The application includes multiple conclusory assertions that Olelo's quality and accuracy are "comparable" to CA-assisted service quality and accuracy.[57] But these claims are neither quantified nor substantiated, and they constitute subjective *qualitative* assessments rather than *quantitative* assessments. And given that the company has provided no evidence regarding Olelo's performance, the

[55] *Id.* at 7.

[56] Various other claims made by MachineGenius—including the repeated assertion that its service is scalable and cost-efficient, *see id.* at 4-5, 7, 10, Ex. B at 1, 2—are also insufficiently explained and appear to be untested. The company provides no cost data or financial information to support its efficiency claims, and it provides no information on how it will scale or its ability to perform at scale.

[57] *See, e.g.*, *id.* at 7 (claiming that the Olelo "captioning accuracy is comparable to the accuracy provided by CAs").

Commission is unable to evaluate these claims for accuracy. In any event, MachineGenius acknowledges that ASR is "*approaching* human-level transcription."[58] By conceding that ASR-only service is merely "approaching" the quality of CA-assisted IP CTS, MachineGenius seemingly acknowledges that ASR-only service has not reached the point of functional equivalence.

MachineGenius also acknowledges that **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**, which again amount to

[58] *Id.* (emphasis added).

[59] *Id.*. Ex. B at 1.

[60] *Id.* at 1-2 (emphasis added).

[61] *Id.* at 2.

[62] *Id.* at 1.

[63] *Id.* at 2.

subjective and qualitative assessments rather than objective quantitative comparisons akin to those demanded by the *Declaratory Ruling*.

Similarly, MachineGenius's operational and technical standards require **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]** Indeed, MachineGenius seeks certification on the theory that functional equivalence may potentially be achieved at some point in the future in the event that ASR technology advances beyond its present capabilities.

MachineGenius's application was filed before the Commission adopted the *Declaratory Ruling*. MachineGenius had the opportunity to resubmit its application to bring that filing into compliance with the testing requirements and guidance provided by the Commission, but declined to do so. MachineGenius's currently pending application should be denied so that it has the opportunity to reevaluate its analysis and bring its filing into compliance with the rules and clarifications provided by the Commission.

In addition, MachineGenius does not demonstrate that Olelo will be offered in a manner that is consistent with the Commission's rules and protects user privacy and confidentiality. To the contrary, MachineGenius concedes that **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted]

[64] MachineGenius App., Compliance Plan Exhibit at 5.

[redacted]

65 MachineGenius App. at 9 n.11.

66 *Id.* Ex. B at 3.

67 *Id.* at 3; *see also id.*, Compliance Plan Exhibit at 6 (similar).

68 MachineGenius App. Ex. B at 3.

69 *Declaratory Ruling*, 33 FCC Rcd at 5834 ¶ 63; *see* 47 C.F.R. § 64.604(a)(2); *see also* 47 U.S.C. § 225(d)(1)(F) (requiring the Commission to promulgate regulations that "prohibit relay operators from . . . keeping records of the content of any such conversation beyond the duration of the call").

70 *See, e.g.*, Geoffrey A. Fowler, *Alexa Has Been Eavesdropping on You This Whole Time*, Wash. Post (May 6, 2019), https://www.washingtonpost.com/technology/2019/05/06/alexa-has-been-eavesdropping-you-this-whole-time/ (discussing "eavesdropping" by Alexa, Amazon's personal assistant, and the retention and use of Alexa recordings by Amazon employees); David Gilbert, *Facebook Said It Wasn't Listening to Your Conversations. It Was.*, Vice (Aug. 14, 2019), https://www.vice.com/en_us/article/wjw889/facebook-said-it-wasnt-listening-to-your-conversations-it-was (discussing Facebook's recent admission that it employed "third party contractors to transcribe the audio messages that users exchanged on its Messenger app"); *see also* Kevin Granville, *Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens*, N.Y. Times (Mar. 19, 2018), https://www.

[redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**[71]

MachineGenius's application also fails to demonstrate that Olelo would be adequate for 911 and other emergency calls. As a mobile-only, ASR-only service, to the extent that Olelo also contracts for or provides the underlying telecommunications routing to the public switched telephone network, it raises concerns with respect to 911 call routing and E911 location accuracy[72]—as well as captioning quality when conditions are not amenable to ASR performance (*e.g.*, background noise, stress levels of the speakers, etc.). MachineGenius offers only vague assertions as to how emergency calls are supported by Olelo. **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]** Moreover, MachineGenius provides no information about its failover plans in the event that circumstances during a 911 or

nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained html (discussing harvesting of personal data from Facebook profiles by Cambridge Analytica without users' consent).

[71] MachineGenius suggests that its privacy policies are permissible because "[n]o part of the MachineGenius IP CTS service will keep any records of the content of any conversation beyond the duration of the call" (MachineGenius App. at 13) and because **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]** (*id.* at 13). But it is not clear that this latter representation satisfies the requirements set forth in the mandatory minimum standards.

[72] *See In re Implementing Kari's Law and Section 506 of RAY BAUM's Act Inquiry Concerning 911 Access, Routing, and Location in Enterprise Communications Systems*, Report and Order, PS Docket Nos. 18-261, 17-239, GN Docket No. 11-117, FCC 19-76 ¶¶ 210-216 (2019); *id.* at App. A, § 9.14(e).

emergency call hamper its ability to generate accurate, automated captions.[73] Such difficulties may be more likely to arise during emergency calls because of the unusual speech patterns and greater background noise that can be expected on such calls.[74]

Finally, much of MachineGenius's application focuses on the costs rather than the service, arguing that **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]** MachineGenius—which argues at length that its program will result in cost savings—fails to appreciate that the question whether reductions in cost outweigh disparity in performance is relevant if and only if MachineGenius can establish that ASR delivers functionally equivalent performance. As the D.C. Circuit has confirmed, consistent with the ADA, the Commission's "primary objective" under Section 225 must be to ensure that individuals with speech and hearing impairments have access to effective communications by telephone; it may consider cost savings or efficiency only to choose between two equally effective alternatives of providing functional equivalence.[76]

[73] *See generally* MachineGenius App. at 16-17.

[74] *See supra* note 37 and accompanying text.

[75] *Id.* Ex. B at 1 (emphasis added).

[76] *See Sorenson Commc'ns, LLC v. FCC*, 897 F.3d 214, 227-28 (D.C. Cir. 2018); Comments of CaptionCall, LLC, CG Docket Nos. 13-24, 03-123, at 22-23 & n. 71 (Sept. 17, 2018).

B. VTCSecure, LLC

VTC is a Florida-based "software communications company that focuses on state-of-the-art communication access for people with disabilities."[77] VTC reports that it employs **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**[78] VTC's proposed IP CTS offering is "VTCSecure App," a free over-the-top app that will be available on mobile devices and desktops.[79] The VTCSecure App will provide ASR-only IP CTS as an option for users.[80] According to VTC, its IP CTS offering reflects a number of "significant improvements in IP CTS technology," including the ability to use video while on an IP CTS call, the ability to use "HD voice," and support for T.140 real time text.[81] Although VTC's proposed offering differs in some respects from that of MachineGenius, it too has not met its burden of showing that it satisfies the Commission's requirements.

VTC's application—like that of MachineGenius—reflects the fact that the company's ASR offering has not yet been subjected to adequate testing. VTC claims that, "[u]nder ideal conditions, VTCSecure has seen over 99% accuracy in situations where there is no Communications Assistant ('CA') and the ASR engine is receiving HD voice."[82] Similarly,

---

[77] VTCSecure, LLC Internet-Based TRS Certification Application, CG Docket 03-123, at 1 (May 26, 2017) ("VTC App.").

[78] *Id.* at 25.

[79] *Id.* at 1, 9.

[80] *Id.*

[81] *Id.* at 1; *see also id.* at 3. It is unclear to what extent these "advancements" will improve an IP CTS user's experience. Although the use of video will enable customers to lip read (*id.* at 2), that option is of limited utility given that a person cannot both lip read and process live captioning at the same time.

[82] *Id.* at 2.

VTC suggests that the "use of ASR technology allows for the potential elimination of the CA for a IP CTS calls [sic] in perfect conditions."[83] But the claim that ASR technology can achieve 99% accuracy under "ideal" or "perfect" conditions is meaningless without further explanation of what those conditions are (or how often they occur)—which VTC never provides. As VTC concedes, ASR **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]** Moreover, the fact that 99% accuracy can be achieved when "the ASR engine is receiving HD voice" is also meaningless given that VTC does not claim that HD voice is always (or even frequently) available to users. Moreover, VTC provides no information at all about what testing, if any, the company has performed to substantiate the figures and qualitative assessments provided in the application. The Commission therefore cannot verify or examine the testing protocol or the test results, and has no assurances that ASR will provide a functionally equivalent experience.

Similarly, VTC's claim that "[a]dvancements in ASR allow it to be extremely accurate and almost on par with human speech recognition"[85] is concerning. The suggestion that ASR is "almost" on par with human speech recognition is essentially an admission that ASR-only service will not be functionally equivalent. But in any event, this claim—made with citation to

[83] *Id.* at 4.

[84] *Id.* at 3.

[85] *Id.*

an article concerning the use of "[d]eep [l]earning" technologies at IBM[86]—does not appear to relate specifically to VTC's IP CTS offering. Indeed, it is not even clear from the application what technology will power the VTCSecure app; VTC says only that an unspecified "remote Artificial Intelligen[ce] system" will "convert the voice almost instantly into text."[87] The Commission does not know whether that technology is endogenous, is somehow affiliated with the IBM technology mentioned in the application, or **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]** If the technology is endogenous, VTC provides no information what the technology is, how it was developed and tested, or whether it is comparable to the best available exogenous technologies (which themselves may not be sufficient to deliver functionally equivalent service). If the technology is exogenous, then it is remarkable that VTC **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**

VTC states that it will deliver a "hybrid" ASR service **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted]

86 *Id.* at 3 n.4.

87 *Id.* at 3.

88 *Id.* at 4.

[redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**

VTC has not shown that its proposed hybrid service meets the standards in the *Declaratory Ruling*. *First*, **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**

89 *Id.* at 3-4.

90 *See supra* note 12.

91 VTC App. at 7.

92 *Id.* at 25.

93 *Cf.* 47 C.F.R. § 64.604(a)(1)(i) (noting that IP CTS providers "are responsible for requiring that all CAs be sufficiently trained to effectively meet the specialized communications needs of individuals with hearing and speech disabilities").

Second, VTC indicates that it plans to have CAs available for certain types of calls, like emergency calls.[94] But to confirm that its ability to rely on CAs is not purely conceptual, VTC should provide additional information and testing data, which should demonstrate that—whichever modality is used to generate captions—the VTCSecure App will "afford a level of quality that is at least comparable to currently available CA-assisted IP CTS with respect to captioning transcription, delays, accuracy, speed, and readability."[95]

C. Clarity Products, LLC

Clarity is a privately held Tennessee corporation that has historically specialized in the production and sale of amplified telephones and other assistive listening devices.[96] Clarity's proposed IP CTS offering is an app called "CaptionMate" that will generate transcriptions of calls and make those transcriptions available for viewing on phones, tablets, or a website.[97] Clarity—**[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted]

[94] *See* VTCSecure Waiver at 3 ("If for some reason ASR is being used during an Emergency call we would like a waiver for the need to present identification to the agent. However, VTCSecure's plan is to always have an agent come on during an emergency call. For the sake of time, that call may start using ASR with the agent coming on the call at any point after the call has started."). This description is quite difficult to follow—creating questions as to whether, and under what circumstances, VTC will start an emergency call with a CA, and raising concerns about VTC's ability to handle 911 and other emergency calls.

VTC's recently filed waiver also potentially raises new privacy concerns. VTC requests a waiver of the Commission's confidentiality rule "where only ASR is used, . . . [s]ince during ASR calls there is no CA to alter the conversation." VTCSecure Waiver at 2. This phrasing is unclear, but granting a waiver of 47 C.F.R. § 64.604(a)(2) without limitation could be construed to permit VTC to record call content to improve its ASR performance, which would be inconsistent with the Commission's commitments in the *Declaratory Ruling*, discussed above.

[95] *Declaratory Ruling*, 33 FCC Rcd at 5834 ¶ 63; *see also supra* note 23. While the Commission was (appropriately) flexible in allowing providers to choose different methods of integrating ASR into hybrid services, *see id.* at 5833 ¶¶ 61-62, it should require and review testing data to confirm that the provider's method does not result in any degradations of service quality for the user.

[96] Clarity App. at 2.

[97] *Id.* at 4-5.

[redacted]**:END HIGHLY CONFIDENTIAL INFORMATION]**—proposes to use an ASR-only system for IP CTS.[99]

Clarity's application does not show that the CaptionMate service has been subject to adequate testing. The application explains that Clarity's testing to date has been limited to a handful of "internal" and "in-house" tests, performed by the firm itself and not by a third party.[100] The application does not identify the call scenarios or the application protocol. Troublingly, what few test results Clarity has provided suggest that there are call scenarios where CaptionMate will not perform well. For example, the application notes that the quality of CaptionMate's transcription is "dependent on the clarity of the remote speaker's voice as well as the amount of background noise."[101] As to other issues critical to functional equivalence—such as CaptionMate's ability to transcribe accented voices and accurately render punctuation, abbreviations, tense, plurals, lingo, jargon, slang—the application provides no information at all.

As a result, Clarity has not demonstrated that it meets the *Declaratory Ruling*'s requirement that applications "should support all claims regarding the[] use of ASR" with "documentary and other evidence" such as "trials and quantitative test results" confirming the ability to match existing service quality levels.[102] For example, Clarity claims that "[i]nternal testing has shown a very high level of accuracy" and a "very quick response rate."[103] But terms

98 *Id.* at B-1 to B-3.

99 *Id.* at 6.

100 *Id.* at 6, D-1.

101 *Id.* at C-1.

102 *Declaratory Ruling*, 33 FCC Rcd at 5834 ¶ 63.

103 Clarity App. at 6.

like "very high" and "very quick" are vague, subjective, and *qualitative* results rather than the type of *quantitative* results contemplated by the *Declaratory Ruling*. The claim that CaptionMate has a "very high level of accuracy" **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**[105] The limited, in-house testing conducted by Clarity therefore does not satisfy the standards articulated in the *Declaratory Ruling*.[106]

Clarity's response, which relies almost exclusively on the 2016 MITRE tests to claim that its service should be certified, does not address these concerns.[107] As discussed above, under the *Declaratory Ruling*, this is insufficient. As an initial matter, Clarity does not suggest that the particular ASR technologies studied by MITRE are the same ones that power its CaptionMate product, and thus the MITRE studies are irrelevant to the question whether Clarity specifically can satisfy the Commission's mandatory minimum standards or CaptionMate specifically can deliver functionally equivalent service. In any event, as discussed above, the MITRE studies did

---

[104] *See, e.g., id.* at C-1. The application lacks details about the accuracy of the CaptionMate service. The application says that CaptionMate "uses contextual clues to correct itself and provide the most accurate transcription possible" (*id.* at 6), but it does not definitively say that the Commission's accuracy standards can be met using ASR only. Phrased differently, the fact that the CaptionMate is as "accurate [as] possible" *using ASR only* does not mean that the transcription will be as accurate as is needed to satisfy the Commission's rules.

[105] *Id.* at D-1 to D-2.

[106] *See supra* note 24.

[107] *See Clarity May 30 Letter* at 2

not simulate real-world conditions and thus are insufficient to support the certification of a provider to offer compensable service—as the Commission itself acknowledged.[108]

Moreover, Clarity's plan to rely on customer feedback provided through a "star"-type rating system does not adequately demonstrate CaptionMate's quality will match the quality of current CA-assisted IP CTS, as is required by the *Declaratory Ruling*.[109] As explained in the application, the star system will ask users to rate calls from 1 to 5.[110] **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**[112]

The application also does not meet its burden of showing that CaptionMate would be adequate for 911 and other emergency calls. As an app to run on phones and tablets, it is likely that Clarity itself is providing or contracting for the routing for calls to the public switched

[108] *See supra* note 10 and accompanying text. Clarity's arguments regarding testing are also discussed above. *See supra* notes 26-27 and accompanying text.

[109] *See Declaratory Ruling*, 33 FCC Rcd at 5834 ¶ 63.

[110] Clarity App. at 5.

[111] *Id.* at 13.

[112] *Id.* at 13-14, B-2.

telephone network, and thus must meet the Commission's 911 and E911 requirements.[113] Clarity's current "Pre FCC" webpage says that "EMERGENCY CALLING IS NOT AVAILABLE during [its] beta test period," and warns users, "DO NOT use this application for dialing 911."[114] Clarity must explain how it will route 911 calls to the appropriate PSAP and provide the required location and callback information. And Clarity's plan for providing functional equivalence for captioning 911 calls is not at all clear from its application. Clarity's application does not describe an adequate failover strategy for providing accurate captions when its ASR engine may be struggling. Because Clarity does not currently or propose to employ any CAs, it is not clear how—if it all—Clarity would be able to protect customers at a mission-critical time, when they are facing an emergency but are experiencing difficulties or malfunctioning with captioning.

Finally, Clarity's application fails to demonstrate that CaptionMate will be offered in a manner that satisfies the Commission's confidentiality rule for TRS.[115] As noted, the Commission specifically asked applicants to provide information about the measures they will take "to ensure compliance" with that rule.[116] CaptionMate **[BEGIN HIGHLY CONFIDENTIAL INFORMATION:** [redacted]

[113] *See supra* note 72.

[114] *Registration*, CaptionMate, https://captionmate.com/ (last visited Sept. 22, 2019).

[115] *See* 47 C.F.R. § 64.604(a)(2).

[116] *Declaratory Ruling*, 33 FCC Rcd at 5834 ¶ 63.

[117] Clarity App. at C-1; *see id.* at C-10 to C-12, E-1.

[redacted] **:END HIGHLY CONFIDENTIAL INFORMATION]**[119] The Commission should confirm that Clarity will satisfy the standard set forth in the Commission's rules.

CONCLUSION

Based on the foregoing, CaptionCall urges the Commission to adopt a framework for evaluating ASR-based IP CTS applications. Even under the existing standards, however, none of the pending applications for certification to provide ASR-based IP CTS have met their burden of proof and thus they should not be granted at this time.

Respectfully submitted,

/s/ Rebekah P. Goodheart

Rebekah P. Goodheart
Elliot S. Tarloff
James T. Dawson
JENNER & BLOCK LLP
1099 New York Avenue, NW
Suite 900
Washington, DC 20001
(202) 639-6000

Counsel for CaptionCall, LLC

September 25, 2019

[118] *Id.* at C-3.

[119] *Id.* at E-1.

APPENDIX A

COMMENTS ON POTENTIAL FCC ACTIONS CONCERNING THE INTRODUCTION OF IP CTS USING ONLY AUTOMATIC SPEECH RECOGNITION TRANSCRIPTION

Richard M. Stern
Carnegie Mellon University
September 25, 2019

I. INTRODUCTION

My name is Richard Stern, and I am a Professor in the Department of Electrical and Computer Engineering, the Language Technologies Institute, and the Department of Computer Science at Carnegie Mellon University (CMU). I have been working in the development of automatic speech recognition systems since 1982, and I am particularly well known for my work in developing novel signal processing techniques to improve speech recognition accuracy in difficult acoustical environments. Among other things, I served on and chaired a number of the committees for the Defense Advanced Projects Agency (DARPA) in the 1990s that established the standards used to define the performance of speech recognition systems. Through these experiences I obtained extensive first-hand knowledge of the development of objective evaluation standards for speech recognition systems. A more verbose summary of my professional activities may be found in the Appendix, and my complete Curriculum Vitae is provided as Exhibit 1.

I have been engaged by CaptionCall to review the state of the art of automatic speech recognition (or ASR, also sometimes referred to as speech-to-text or STT) technology as it applies to Internet Protocol Captioned Telephone Service (IP CTS) and to opine on the feasibility of the exclusive use of ASR technology for transcription of telephone calls by IP CTS providers. To prepare this report, I reviewed the materials that are cited herein and the materials before the Federal Communications Commission regarding the state of ASR technology, including studies

performed by MITRE Corporation. My conclusions are based on these materials and my experiences working in this field for several decades.

My principal conclusion is that while the capabilities and accuracy of ASR systems have improved dramatically in recent years, there is insufficient evidence that the current ASR technology is at a level that is necessary for use in IP CTS without the real-time participation of human communications assistants (CAs).[1] I further conclude that the premature implementation of ASR-only IP CTS at the present time is likely to degrade significantly the usability and effectiveness of these services for hearing-impaired users.

My prime reason for concern about the provision of ASR-only IP CTS is that there is a great deal of variability in the nature of the speech signal. Perhaps most critically, despite the extraordinary advances in ASR performance for baseline conditions, I have not seen any data, test results, or evidence indicating that ASR-only providers have overcome the well-documented problems that arise with respect to the speech of multiple classes of speakers including women, the elderly, young children, members of minority groups and individuals who speak in dialects, non-native speakers, speakers under stress, hearing-impaired speakers, and individuals suffering from various types of neurological impairment including Parkinson's disease, cerebral palsy, etc. As many have noted, these communities are already suffering from multiple other societal disadvantages and the implementation of IP CTS services based on ASR alone will inevitably further disenfranchise them. In addition, the performance of ASR systems is also degraded by distortions imposed by the background acoustical environment. I have not seen anything to suggest or demonstrate that these problems can be fully compensated for—and such distortions

[1] My conclusions do not address the possibility of integrating ASR into a "hybrid" service—for example, where an IP CTS provider can identify call types, perhaps IVR or business-to-business calls, for which the variations that present difficulties for ASR technology may be less prevalent.

are likely to be common in many live telephone call scenarios. Finally, and importantly, present ASR technology cannot determine reliably when a new word is spoken that is not already part of the system's internal vocabulary.

Nevertheless, there is room for optimism about the evolution of ASR technology. Approaches to compensate for all of these types of variabilities are under development around the world, including in my own research group. Eventually the major ASR service providers will develop products that serve all individuals much more equally, both because it is the right thing to do and because it will increase the number of potential customers. While I do not believe that present ASR technology is sufficiently robust to serve as the sole transcription modality for IP CTS, I have every reason to expect that the technology will attain a satisfactory level of performance for all populations of potential users in the future. At that point, the FCC and IP CTS community will need to have procedures in place to evaluate potential IP CTS providers on a case-by-case basis and to determine whether their systems are suitable for IP CTS. I provide some suggestions toward this end at the conclusion of this report based on the successes that I witnessed firsthand while working through the processes established by DARPA and the National Institute of Standards and Technology (NIST).

This report is organized in the following fashion: I begin with a brief description of human speech production and current ASR technology. This is followed by a detailed discussion of the various major sources of speech variability and the difficulties that they pose for ASR systems, along with a less detailed discussion of the impact of acoustical variability and the important problem of "new" words that are not known by the system. The final section of this report discusses the evaluation of ASR systems. I begin with a brief summary of the history of the very successful evaluation program for ASR systems put in place by DARPA and NIST. I then

comment on the recent evaluation of ASR for IP CTS conducted by the MITRE Corporation, which appears to have had some influence on the FCC's decision to solicit comments on the feasibility of ASR-only IP CTS. Finally, I make some suggestions concerning how the DARPA/NIST experiences may be applied to the evaluation and of ASR-only IP CTS once technology has improved to the point where it is ready for introduction into the market.

II. BACKGROUND AND OVERVIEW OF AUTOMATIC SPEECH RECOGNITION TECHNOLOGY

A. Provision of IP CTS

To evaluate issues associated with the provision of IP CTS using ASR technologies, I have learned about the current provision of service using CAs and about the current user base by reviewing relevant materials from the FCC's proceeding involving its IP CTS program and by discussing those materials with CaptionCall's legal counsel.

It is my understanding that IP CTS is a type of telecommunications relay service (TRS), which is specifically authorized under the Americans with Disabilities Act to enable individuals with hearing and speech impairments to communicate effectively by telephone. As I understand it, IP CTS is particularly effective for individuals with hearing loss but who still have some residual hearing, and communicate most effectively by relying on their remaining hearing and speech, while also reading captions of what the other party on a telephone call says generated by an IP CTS provider.

It is my understanding that the primary model for providing IP CTS is currently by having a CA hear what the non-user on a telephone call says and revoice that speech into speech recognition software that is specifically trained to that CA's voice. (I understand that one provider uses stenographers, rather than the CA-revoicing model.) The speech recognition software then generates captions, which are provided to the user over an internet connection that is separate from

the call stream. Under this current model, the CA sees the captions that are displayed for the user, and is able to make corrections as necessary, which are highlighted for the end user. The CA does not otherwise communicate with the user or the non-user on the telephone call.

I further understand that IP CTS providers must be "certified" by the FCC to receive compensation for providing service to eligible users. The FCC has not yet developed service quality metrics, testing data sets, test methodologies, evaluation protocols, or mandatory standards that it can use to evaluate applicants to provide IP CTS—regardless of whether the applicant intends to use CAs or only ASR. I understand, however, that the Commission has contracted with MITRE to study the performance of IP CTS providers, including ASR-based providers. And I further understand that the industry of IP CTS providers is currently working toward the development of metrics, testing datasets and methodologies, and, ultimately, service quality standards that could apply to all IP CTS providers. I am not working with the industry group, but I have reviewed their two submissions to the FCC.

B. Production of Human Speech

It is worthwhile to begin by briefly reviewing how human speech is produced. A useful model for speech production is a sound source that has its frequency characteristics shaped over time by a filter [1, 2]. The sound source in the case of speech production is either a periodic train of pulses (produced by passing air from the lungs through the vocal chords) or broadband noise (produced by passing a stream of turbulent air directly through the throat without interference by the vocal chords). The fundamental frequency of the periodic pulses can be manipulated by increasing or decreasing the air pressure from the lungs (for example, when we hum a melody: The resonant frequencies of the vocal tract are determined by changes in the cross-sectional area along its length, which we manipulate as we produce different vowels. For example, in producing the sound "ah" we widen the part of the tract close to the mouth; in producing the sound "ee" we

close the mouth but widen the part of the tract near the back of the throat. The vowels that we perceive are determined by these resonant frequencies, called *formant frequencies*. Consonants are produced by briefly stopping the flow of air by holding the tongue against the lips (as in producing "ba" and "pa"), the teeth ("da" and "ta"), or the roof of the mouth ("ga" and "ka"), among other locations. We perform all of these acrobatic manipulations of the articulators that produce speech sounds subconsciously. We learn to produce these sounds, and recognize, and discriminate among, them by listening to and mimicking the sounds that we are exposed to in early childhood.

The interaction between the fundamental frequencies produced by the periodic pulses of the vocal chords and the formant frequencies representing vowels that are formed by the throat and articulators is one of the primary reasons why speech produced by women and children is often more difficult for ASR systems to recognize accurately than speech produced by men.

C. Basic Speech Recognition Technology

Automatic speech recognition is essentially a special class of *pattern classification* algorithms that guess which of a number of possible "classes" of input is actually present. All pattern classification systems operate on the same basic principles: an initial analysis stage performs a physical measurement (of a sound pressure wave, in the case of speech recognition) and transforms that measurement into a set of *features*, or numbers that are believed to be most indicative of the classification task to be performed. These features typically indicate which input class is present, but they are corrupted by a degree of random variation (both because of variability in the way the speech signal is produced, and because of variability in the acoustical environment). A second component decides which of the possible inputs is most likely, based on the observed values of the features. The features that are extracted typically describe (directly or indirectly) the instantaneous *frequency response*, or relative power of the various frequency components during

each brief observation interval. These features are also motivated to a greater or lesser degree by how humans produce and perceive speech.

Speech Recognition Based on Hidden Markov Modeling. The technologies for determining the most likely word sequence from a spoken utterance have evolved greatly over the decades. From the early 1980s until very recently the dominant speech recognition technology has been the *hidden Markov model* (HMM, [1, 3]), and practical systems based on HMMs remain in widespread use today. The HMM representation characterizes the incoming speech waveform as an unseen time-varying process that is random (or stochastic) in two ways: there is assumed to be a hidden random sequence of states that correspond to the underlying phonemes (which are the fundamental building blocks of speech sounds), and each state transition is represented by a set of observable random numbers that are compared to the incoming features. The task of the recognition system is to infer the identity of the unobserved state transitions (and hence the sequence of phonemes) from the observed values of the features.

The performance of an HMM system depends most critically on the accuracy of the probabilistic model that characterizes how the observations from the state transitions depend on the phoneme sequence that is input to the system [1]. HMM systems today most commonly use phonetic models that are linear combinations of Gaussian densities, or *Gaussian mixture models* (GMMs). HMMs using Gaussian mixtures for the phonetic models are frequently referred to as "HMM-GMM systems."

The technologies for training HMM systems efficiently and accurately have evolved greatly over decades. While a detailed understanding of these training systems is not necessary for present purposes, at a high level these systems attempt to estimate the form of the mathematical

models that describe the observations and the state transitions using large numbers of labelled examples of spoken word sequences.

Speech Recognition Using Deep Learning. While the HMM-GMM paradigm remained the dominant speech recognition technology from the early 1980s through the mid-2000s, new approaches to speech recognition based on *deep learning* are becoming more popular. The structures that implement deep learning are frequently referred to as *artificial neural networks* or *computational neural networks*.

While the basic approaches to pattern classification using computational neural networks have been known for some time, these approaches have become more effective and practical in recent years because of a better understanding of the capabilities of the underlying mathematics, the widespread availability of much larger databases for training, and much faster computing infrastructure, including the availability of *graphics processing units* (GPUs), which are particularly well suited for many of the core computations associated with neural networks.

In general, speech recognition systems based on deep learning have the advantage that the all-important model of the feature values produced by a particular phoneme sequence may be of any form, rather than limited to a pre-determined parametric form such as the GMM. They have the disadvantage of requiring substantially more training data than conventional HMM-GMM systems because the model that represents the observations is no longer assumed to take on a particular form (like a Gaussian). In addition, it is widely observed that the recognition accuracy of DNN-based ASR systems degrades more rapidly than the accuracy for HMM-GMM-based ASR systems when speech is input that is different from the training data.

While neural networks were initially used to produce better phonetic models in a system that incorporated a traditional HMM for the decoding component (*e.g.*, [4]), other architectures are

becoming more popular in which the entire end-to-end speech recognition process is performed using a chain of deep neural networks (*e.g.*, [5]). ASR systems based on deep learning are increasingly utilized because they provide consistently better acoustic-phonetic models than the traditional Gaussian mixtures for baseline conditions; indeed, they have undergone explosive growth and development in recent years (*e.g.*, [6, 7]).

III. LIMITATIONS ON THE USE OF ASR IN IP CTS

Based on my background in speech recognition and my review of the materials described above and cited herein, I was asked to evaluate whether ASR technology is currently capable of being used to provide IP CTS without human involvement. As I explain, I believe that the introduction of ASR-based IP CTS, without human involvement, would expose users to a risk of degraded service quality and could introduce new forms of discrimination into the provision of service. As I explain below, I have not seen any evidence, data, or information suggesting that companies have identified or developed solutions to well-documented problems that will likely render ASR-only IP CTS less accurate and effective than CA-provided IP CTS.

A. Challenges for ASR Systems in IP CTS Applications

In this section, I discuss a number of the classic obstacles that presently would hamper the use of ASR to provide IP CTS. These obstacles include speaker-based variability, environment-based variability, the new word problem, and training issues for these systems. These problems, which apply to any use of ASR, are likely to be particularly difficult to overcome for ASR's use in IP CTS, because the recognition system must serve an essentially unlimited variety of speakers (*i.e.*, anyone with whom the IP CTS user may speak with on the telephone), each with his/her own idiosyncratic ways of pronouncing words. These speakers could be anywhere during the call such as in his/her home, a car, a city street, a park, a stadium, an airport or train station, a hospital, and so forth.

Let us begin by considering a few types of user experiences that are likely to be difficult for ASR-only IP CTS but that may be relatively common use cases for the service:

- A consumer with hearing loss calls an emergency services call center using a 911 service concerning symptoms she is experiencing that might indicate a heart attack. The emergency responder is trained to avoid having stress modulate his speech patterns, but there is a great deal of ambient noise in the call center which makes it more difficult for the service provider to interpret what the caller is saying and respond appropriately. The background noise includes the speech of other responders at the call center, announcements over the center's PA system, and ambient noise from equipment at the facility. The diagnosis of the user's problem potentially involves some specific unusual procedures and queries.
- An individual who is older has aged into hearing loss. He maintains contact with this family primarily through phone calls. These calls involve multiple handoffs between his daughter, son-in-law, and two young grandchildren, and sometimes occur while the family is in its home; sometimes, when the family is at a playground or outdoors. The individual's support team worries that without these phone calls, he might experience social isolation and depression.
- A social worker who is beginning to experience hearing loss must communicate with multiple clients by telephone, given the limitations on her time and resources. She works with individuals with different English language skills, some of whom speak non-standardized dialects, some of whom have pronounced accents, and some of whom have cognitive or related disabilities. Regular contact with her clients is essential, and sometimes involves discussions of treatment protocols. One of her clients is taking a

course of antiretroviral therapies for HIV, which must be taken at precise times and in precise doses.

- It is my understanding that individuals use IP CTS to talk to their lawyers, doctors, financial advisors, and other service providers. These conversations are likely to involve unique vocabularies and specialized acronyms that are necessary to convey detailed information, including information which may be highly sensitive to the user.

As the above examples illustrate, telephone calls involve multiple axes of variability that will present problems for ASR-only IP CTS, which I discuss below in the context of existing research on speech recognition. Based on my experience, I believe that it is likely that permutations of speakers, speech content, and background conditions can have a compounding adverse effect on the ability of an ASR-only IP CTS system to generate accurate captions on which a user can rely.

Speaker-Based Variability. Perhaps the most important issue for this discussion is variability with respect to speaker, and the consequent high likelihood that multiple high-risk populations would be denied access to IP CTS of equal quality and usability. Some of the common sources of speaker-based variability that are most relevant to the quality of IP CTS include the following:

- **Gender.** From the earliest days of research into ASR technology, it has been widely observed that recognition systems perform better for male speakers than for female speakers (*e.g.*, [8, 9]) for multiple reasons. Most women produce speech with a higher fundamental frequency than most men, and the interpretation of this speech is confounded by interactions between the pitch frequencies and the formant frequencies described above [10]. Furthermore, many databases used to train ASR systems that are based on

purportedly naturally-occurring "found speech" tend to have more examples of male speech than female speech. In addition, there has been some speculation that when men speak spontaneously, they may speak more aggressively—and, for ASR systems, more clearly—than women.

- **Age.** A number of studies have considered how ASR accuracy varies as a function of age (*e.g.*, [11]). In general, recognition accuracy is worse for children below the age of 15 and individuals who are greater than about 70 years old [12]. The speech of young children is highly variable, which increases the likelihood of a mismatch between training and testing conditions for ASR systems. In addition, ASR performance suffers from similar difficulties with children as with women, although to an even greater extent, as their speech is very high pitched. Increased ASR error rates for people who are older are attributable to multiple factors including changes in the respiratory system, the larynx, and in the vocal tract, combined with a decrease in neuromuscular control. Jitter, shimmer, and breathiness are introduced into speech produced by people who are older [13]. All of these factors contribute to mismatches between training and testing conditions. Independently of mismatches, the greater variability of the speech produced by people who are older renders their speech intrinsically more difficult to recognize.
- **Race and Dialect.** Variations in dialect also produce substantial increases in ASR error rate (*e.g.*, [14]), again primarily because of mismatches between training and testing conditions. From a linguistics standpoint, no dialect is inherently more or less intelligible. The main factor which determines how well a listener understands a dialect is the amount of exposure they have had to it; with sufficient exposure, any human listener can learn any language variety [15, 16]. Nevertheless, variations in dialect make a tremendous difference

in how well ASR systems will perform. Unless a system is trained from data that are carefully curated to represent all dialectal variations under consideration, speakers of minority dialects will be at a disadvantage in their ability to use ASR-based services efficiently and effectively. African-American speakers constitute a very important subclass of dialect variability. It has been widely observed that African-American speech tends to be recognized less accurately, both because it typically is not well represented in training data and because of variations in how some words are pronounced. This can have profound implications to the accessibility of modern technology to African Americans and members of other minority groups (*e.g.*, [17]). As a recent study noted, "automatic speech recognition systems continue to struggle to maintain high accuracy in the face of well-documented systematic sociolinguistic variation. This is particularly troubling given that the groups of talkers with the highest error rates, in particular African Americans and talkers who use non-standardized regional varieties, are those who have faced other kinds of discrimination as well [18]."

- **Non-Native Accents.** Unsurprisingly, the accuracy of ASR systems degrades when the speaker is speaking in a second language (*e.g.*, [19, 20]). For example, it is widely observed that the accuracy of English-language ASR systems degrades to some degree when the speakers are of European origin, but much greater degradations in performance are observed when the speakers are of East Asian origin. These degradations occur because non-native speakers produce words less reliably, and because even the same phonemes are rendered differently in different languages—both of which factors produce mismatches between training and testing conditions.

- **Stress.** It is well known that the performance of speech recognition systems degrades under the presence of stress, which in this context refers to speech produced under environmental, emotional, or workload stress (*e.g.*, [21]). Stress conditions also extend to Lombard speech, which refers to speech produced in environments where there is a great deal of noise. Because only a small amount of data is available from individuals who are truly under great stress, ASR systems that are expected to encounter stressed speech must be trained on simulations of stressed speech, which is not always accurate.
- **Hearing Impairment.** Profoundly hearing-impaired individuals typically produce variant speech because these individuals cannot hear and correct the speech that they produce. Up until now there have been very few formal studies of the recognition accuracy obtained from ASR systems with hearing-impaired or deaf individuals. One recent study found that for a particular commercial translating device, ASR error rates for deaf speakers were more than 4 times those obtained for normal-hearing speakers, rendering the device unusable [22]. While this is a small sample, it is expected that these results would be typical.
- **Neurological Disease.** The nature of speech production is also affected by a number of neurological conditions such as dysarthria and Parkinson's disease. Early research investigating the performance of ASR systems on speech from speakers with different degrees of dysarthria such as Friedrichs Ataxia, traumatic brain injury and cerebral palsy, reported in all cases greater word error rates for speech from patients compared to speech from normal controls [23]. All of these conditions create communities of underserved individuals who frequently are also more dependent on social services than average because of their medical conditions or other limitations. The results of a very recent study using a state-of-the-art ASR system confirm that the distributions of word error rates for

sufferers of Parkinson's disease remain consistently greater than those of healthy individuals, although the distributions of scores overlap [23].

It is clear from the discussions above that while the baseline performance of ASR systems has achieved an improved level of accuracy, these systems do not work equally well for all individuals. Articles about this phenomenon in the popular press include "Why some accents don't work on Alexa or Google Home" (The Washington Post [24]), "In the world of voice recognition, not all accents are equal" (The Economist [25]), and "Voice is the next big platform unless you have an accent" (Wired Magazine [26]), among others. The disparities between the ASR accuracy obtained by white North American users and the accuracy obtained by everyone else is real.

Based on the existing literature highlighted above, there is little doubt in my mind that the deployment of ASR-only IP CTS would further marginalize and systematically disadvantage minority users as well as users who primarily speak with individuals who are not white North American males.

Environment-Based Variability. In addition to dealing with extensive potential speaker variability, IP CTS providers must cope with degradation imposed by the acoustical environment. Unsurprisingly, the presence of background noise degrades ASR accuracy. Nevertheless, it is less appreciated that the degree of degradation is highly dependent on the nature of the background interference. For example, background speech or music, or transient interference from sources such industrial machinery impose much more degradation than smoother noise sources such as wind noise or speech babble (*e.g.*, [27]). Reverberation—which would be caused by talkers using speakerphones, talkers in large public spaces, and to some degree in automobiles—is another source of degradation that degrades ASR accuracy. It is more difficult to compensate for the effects of reverberation than for additive noise (*e.g.*, [28]). While research directed to

compensating for these effects has gone on for decades, it is far from completely effective, and the type of compensation needed depends on the nature of the degradation. No single approach handles all of these effects simultaneously (*e.g.*, [29, 28]). A classic study in 1997 observed that "error rates of machines are often more than an order of magnitude greater than those of humans for quiet, wideband, read speech. Machine performance degrades further below that of humans in noise, with channel variability, and for spontaneous speech" [30]. The author went on to suggest that the "human-machine performance gap can be reduced by basic research on improving low-level acoustic-phonetic modeling, on improving robustness with noise and channel variability, and on more accurately modeling spontaneous speech." While some progress has been made in all of the areas cited since that time, substantial gaps between machine and human recognition accuracy remain, especially in degraded conditions.

The "New-Word" Problem. Many studies across all languages have demonstrated that distributions of the frequencies of words are very "heavy tailed" in that if we were to rank order words by their frequencies of occurrence, there would be a large number of words that we use regularly that occur only infrequently (*e.g.*, [31]). Consequently, if an ASR system is trained with a finite vocabulary, there is still a substantial probability that a randomly presented word will be outside the pre-trained vocabulary. This problem is exacerbated when a conversation is about a topic in a particular technical domain such as medicine, law, or engineering (*e.g.*, calls in which patients discuss medications with doctors or calls in which customers seek service for a household device such as a television or refrigerator), which is a very natural occurrence in telephone calls. Historically, it has been very difficult for speech recognition systems to detect when a presented word is outside of the system's pre-trained vocabulary. This problem is addressed in today's commercial systems either by having the system provide implicit or explicit confirmation of what

was recognized (as in conversational systems), or by marketing systems in specialized fields such as medicine or law that are pre-trained with large vocabularies of domain-relevant words. In contrast, IP CTS systems operate in "open loop" fashion, with no opportunity for automated systems to request confirmation or clarification unless an uncertain input is detected. And even when an uncertain input is detected, an ASR-only system does not have a human from whom to request confirmation or clarification.

* * * * *

The problems of speaker and environmental variability that I have discussed are well known in the speech-recognition community, which has attempted to develop techniques to compensate for these issues. But the most effective tools vary from case to case. Critically, most commercial speech recognition systems need to deal—and are designed to deal—with only a limited variety of use cases. For example, command and control systems or voice-based search such as SIRI or Cortana anticipate speech over the telephone in short, well-formed phrases. The Amazon Echo works in homes and responds primarily to simple commands and phrases. All of these systems work best for educated North American users. In contrast, IP CTS providers must be able to cope with the *worst* of all worlds. They must be responsive to speech from any conceivable type of speaker in any conceivable environment with no constraint whatsoever on vocabulary, syntax, or topic under discussion, with a significant likelihood of new words being part of the input. In fact, IP CTS is a *far more difficult task* and operational domain than that of any commercial system that has ever been deployed up to now.

B. Issues Related to ASR Training

The performance of any speech recognition system is only as good as the data with which it is trained, and the quality of the acoustic-phonetic modeling is the most significant determiner of the consequent performance of an ASR system. Because ASR systems can make recognition

decisions only by extrapolating from the data with which they are trained, it is critically important that the systems be trained with a broad range of data that incorporates all of the potential variabilities in speech production that would be encountered when a given ASR system is put into practice. As implied above, for a general-English application like IP CTS, the training data should include sufficient speech from women as well as men, speakers of all ages, non-native speakers, African-American and other minority speakers, and speakers with hearing impairments and neurologically based speech production disorders. It is critically important that the databases used to train ASR systems for IP CTS be balanced in this fashion in order to avoid introducing bias in the performance toward users from privileged groups. To ensure that underrepresented groups are protected from performance bias, ASR systems should be tested to measure accuracy on the various groups individually.

While the baseline performance of modern DNN-based ASR systems is improved, these systems degrade more than the older HMM-GMM systems when they are presented with input that is outside the "envelope" of conditions that were represented in the training data (*e.g.*, [32]). This problem is addressed by the widespread use of "augmented training" procedures that enable DNN-based ASR systems to become more robust with respect to acoustical environments by incorporating a wide variety of simulated acoustical degradations into the training data. Nevertheless, it is never possible to anticipate all of the many types of degradations that could potentially be applied to speech that is input to an IP CTS system.

Some potential ASR-based IP CTS providers may make use of the speech-recognition services offered on a wholesale basis by large companies such as Amazon, Apple, Google, and Microsoft. These organizations have been able to achieve good performance for their ASR products by collecting and retaining speech data provided by users of their services. Since callers

to these companies tend to be articulate, computer-literate adults seeking specific information, the type of conversations and the nature of the speakers are likely to be more limited than the typical conversations and speakers serviced by IP CTS products. This presents a critical problem for the introduction of ASR-only IP CTS products. On the one hand, it is my understanding that IP CTS providers are prohibited from collecting supplementary speech data from their own calls to protect user privacy, which could make it effectively impossible for an IP CTS provider to develop and train its own ASR system. On the other hand, companies like Amazon, Apple, Google, and Microsoft are not subject to such prohibitions, and, if they wholesale their engines to IP CTS providers, they have every incentive to capture as much call content as possible for training purposes—unless they are somehow restricted from doing so. Users of IP CTS may not understand that their calls are being recorded by an underlying ASR wholesale provider, compromising the privacy of their phone calls.

IV. OBJECTIVE EVALUATION OF AUTOMATIC SPEECH RECOGNITION SYSTEMS

I now turn to the objective evaluation of ASR systems. I begin with a review of how the modern evaluation of ASR systems came about. I continue with a discussion of the recent MITRE evaluations of ASR systems, and finally I discuss how the FCC could incorporate best practices from the DARPA/NIST experience to the problem of evaluating and certifying IP CTS providers.

A. Formulations of the Initial Objective ASR Evaluations

The development of collaborative, yet competitive, objective evaluation protocols in the 1980s and 1990s was a critical driver for the development of the ASR capabilities that we all enjoy today. In the 1960s and the 1970s, there had been substantial interest in the development of speech recognition technology, frequently using knowledge-based approaches rather than the statistical

approaches that dominate machine learning systems today.[2] The systems that were developed were typically limited in scope, brittle to changes in input, and incapable of performing at a consistent, high level. Moreover, there were no generally accepted methods for evaluating the performance of ASR technologies, and there was a substantial degree of charlatanism associated with how companies typically described the performance of their products. Many influential people concluded that speech recognition was a hopeless task. Respected skeptics included the highly influential John Pierce, who convinced DARPA to abandon funding for research in speech recognition in the early 1980s [33].

Around 1985, DARPA restarted support for ASR and related technologies in speech and language, which became so successful that it has become the model paradigm for similar technology-development efforts. The most important reason for the success of the new DARPA speech program was that it incorporated a well-defined objective evaluation metric applied by a neutral organization (the National Bureau of Standards (NBS), which later became NIST). The sites participating in the new DARPA speech program (which included my home institution of Carnegie Mellon University and MIT, along with several corporations and research organizations) developed ASR technologies that were applied to a common data set using training data developed by the community. The systems were competitively evaluated once per year based on new data provided by NIST that had not been seen previously, and the results of the evaluation were disclosed within the research community. There was tremendous prestige associated with developing the best-performing system. Each site was expected to disclose its major innovations, and this scenario was repeated in subsequent years with new unseen test data. The nature of the evaluation data, as well as the specifics of the evaluation metrics, were worked out in committees

[2] Some components of this discussion are abstracted from a 2019 presentation by Mark Liberman [34].

that included representatives of all the participating sites, as well as the stakeholder governmental agencies. The evaluation metrics that were developed through this process became widely adopted by the entire research community. For example, my understanding is that the definition of word error rate (WER) developed under the auspices of the new program remains the basis for the measure of WER used to evaluate systems for IP CTS today, with some modification. The software developed by NIST to perform the automatic evaluation was made available to all so that system developers both inside and outside the DARPA community could use it to benchmark progress on an ongoing basis.

The combination of a common evaluation using shared data and an objective evaluation metric was the impetus for decades of unprecedented progress. Figure 1 depicts on a logarithmic axis how WER decreased from the 1980s to the mid-2000s for a variety of speech recognition and understanding tasks. (Once the WER began to approach 5-10 percent, DARPA would switch to a more difficult task.) I believe that the dramatic improvement of ASR during this period was driven by the paradigm that was put in place by DARPA/NIST for the research community.[3] Indeed, this "common task method" has become the standard research paradigm in experimental computation science and has been applied to many other problems including machine translation, speaker identification, language identification, information retrieval, text summarization, video analysis, etc.

As I discuss below in Section C, I believe that, with modifications, this paradigm should be adopted for the evaluation of IP CTS technology. It is my understanding that all of the currently certified IP CTS providers are currently engaged in a similar project that could be used to evaluate

[3] ASR has unquestionably evolved significantly since 2004, and Figure 1 is not included to reflect the WER of current state of the art ASR technology, but rather to illustrate that the FCC can facilitate further innovation by putting in place the right framework.

the performance of CA-based and ASR-only services alike. Conversely, if providers are allowed to develop ASR-only IP CTS based on their own subjective testing, with no transparency into the datasets and evaluation protocols used, and the results obtained, experience teaches that there is a great risk that their services will not perform as advertised.

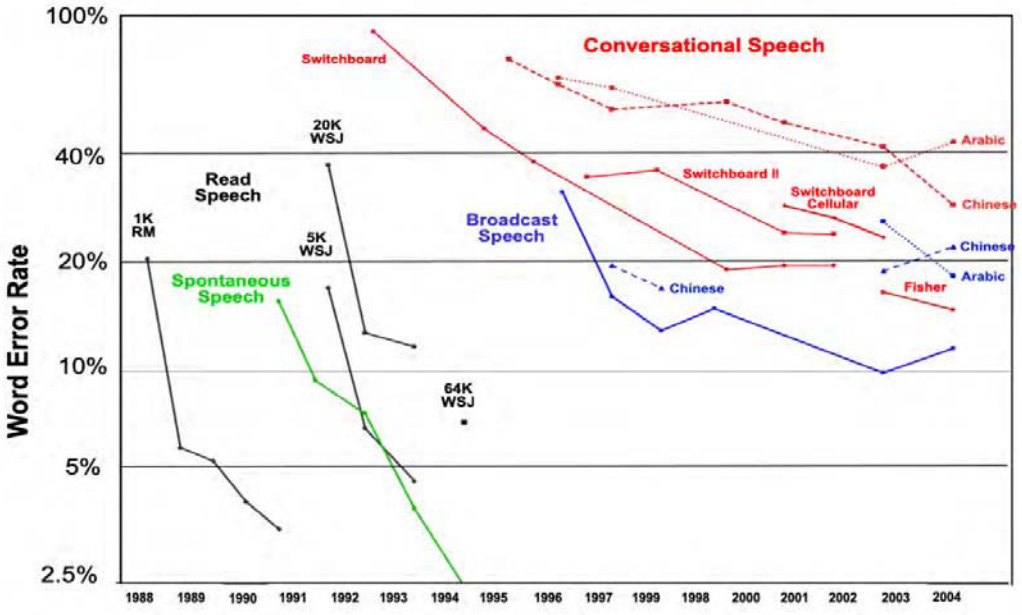

Figure 1. Transcription error for a variety of speech recognition and tasks over years. (From A. Le, NIST)

B. Comments on the Evaluation and Analysis of ASR Systems by MITRE

The MITRE Corporation conducted evaluations of traditional human-based and ASR-only IP CTS systems in 2015-2017, comparing the performance of the systems considered in terms of the delay, accuracy, and usability of the services provided. I focus my discussion on the accuracy statistics reported by MITRE. The Phase I report of the study compared the transcription accuracy of four traditional IP CTS providers and one ASR system. The average accuracy of the four traditional IP CTS providers ranges between 82.8 and 88.7 percent, while the single ASR system had a score of 83.0 percent. The Phase II report of the study compared the accuracy, speed, and usability of four traditional IP CTS providers and three ASR-based IP CTS systems. The overall

accuracy of one of the three ASR-based systems was slightly better than the average of the four traditional providers; the other two ASR systems produced substantially worse accuracy.

While these results may legitimately be taken as an indication that ASR technology could potentially be viable for IP CTS in the long term, these results do ***not*** imply that the ASR technologies are appropriate for immediate introduction to the market, as the MITRE evaluation suggests. For example, in a limited study in the Phase I component of the MITRE evaluation, it was observed that the average error rate increased by a factor of only 2.5 when native speakers were replaced by non-native speakers for traditional IP CTS providers, but the error rate with non-native speakers for the ASR-based systems was more than 7.5 times the baseline error rate for native speakers. Similarly, in another limited study it was observed that the error rate increased by a factor of about 1.55 when white noise at a "moderate" level was added to speech that was transcribed using the traditional IP CTS providers, while the same error rate increased by a factor of 2.12 when an ASR system was used for the transcription. These comparisons indicate that the ASR systems can be more fragile with respect to variability in their input than are the human transcribers.[4] Hence, comparisons of results that are dominated by baseline comparisons are misleading and cannot be extrapolated to the performance under the myriad circumstances discussed above.

Moreover, the measurements of the MITRE study are extremely difficult to interpret clearly because of the lack of information provided about how the ASR systems were trained. There is no specific indication (except for the one study with non-native speakers) about the extent to which speakers used to evaluate the system include a broad range of ages, dialects, education,

[4] Most likely this is a consequence of the much greater ability of humans to acquire, interpret, and integrate information from a wide variety of sources and to continually adapt to new sources of variability such as accent or environmental conditions.

state of health, etc. Similarly, the range of environmental conditions considered in the first two phases of the MITRE evaluations was extremely limited. Although one ASR system in the MITRE studies performed at the levels of the traditional IP CTS providers using human communications assistants, the overall level of results using ASR is substantially worse than the results obtained using traditional IP CTS providers.[5]

All of these considerations compel the reader to respect the actual recommendations of the MITRE study itself: research should continue into the feasibility of fully automated ASR services in place of existing IP CTS services. The MITRE study **does not** recommend the immediate deployment of ASR-based services for the hearing-impaired community.

C. Best Practices from DARPA/NIST for Evaluating and Certifying IP CTS providers

As indicated above, I believe that the present state of the art of automatic speech recognition does not yet justify the introduction of ASR-only IP CTS into the market. A change of service of this magnitude and importance should take place only after an objective and controlled evaluation of each proposed service considering the full range of potential users, as well as an evaluation of each provider-specific service.

---

[5] A third report by MITRE also includes a description of how the Phase III comparative evaluation of IP CTS systems using CAs and ASR would be carried out. Even if MITRE's planned methodologies and design allowed for an apples-to-apples comparison of IP CTS systems, the results would not be generalizable. The size of the study is very small (only 10 users, all of whom must travel to MITRE) and is thus unlikely to be representative of the population of IP CTS users. The results obtained must not be construed to be representative of the accuracy that would have been obtained using speakers from the disadvantaged populations addressed in Sec. II-A of this report. Similarly, there is no systematic consideration of the effects of acoustical environment, nor of the introduction into the evaluation process of "new" words that are not part of the training vocabulary of the ASR system. This means that the MITRE study will assess the performance of ASR for IP CTS under the best possible conditions, but will not address how well ASR will work for the large number of people who do not produce standard North American English speech. My understanding is that the mandate of IP CTS is to serve everyone, not just those individuals who produce the clearest speech.

In addition, I believe that it is extremely important that new scenarios are developed and new speech data are collected each time the qualifying evaluations are performed. The MITRE plans do not make explicit mention of this.

Nevertheless, the ASR technology will clearly continue to improve, and it is an appropriate role for government agencies to create incentives and frameworks to encourage such innovation. It is clear that the present time is ripe for these stakeholders to develop agreed-on standards and evaluation protocols with which future potential ASR-based IP CTS systems can be evaluated. In general, for the FCC to authorize a particular provider, it should understand *that provider's* results in a detailed and transparent fashion. Based on my experience and the general experience of the entire speech recognition community of the 1980s, without an appropriate framework in place, speech-recognition providers are likely to over-promise and under-deliver in terms of their performance. Also, lab results have typically been a poor predictor of field performance. We have often seen systems work well under controlled conditions with well-behaved speakers but fall short when confronted with real customers in a live service. Thus, even if the FCC has reached the general determination that ASR technologies are ready for the provision of IP CTS without human involvement—a determination with which I strongly disagree—it should evaluate the mechanics and processes of each individual ASR engine before reaching a determination as to whether it should be used to provide IP CTS.

I believe that, with modifications that respect the proprietary nature of most technology developed by the IP CTS community, the paradigm developed by the DARPA/NIST speech program in the 1980s should be considered as a model to drive the evolution of ASR technologies to the place where they are ready for use to provide IP CTS. I am optimistic that the technologies have already improved dramatically, and, with the right framework in place, they will eventually be capable of matching CA-based services. This framework should include the following elements:

- A group of providers and, if appropriate, other stakeholders, should establish and define metrics, datasets, and testing methodologies for IP CTS. The group should develop a range of speaker and environmental conditions that would be part of the evaluation for potential providers. This group would function as a fair broker to ensure that IP CTS quality will remain at its present level regardless of whether individual providers use human transcribers, ASR systems, or a combination of the two.
- Testing, performance evaluations, and, ultimately, certification decisions must be based on objective quality. Third-party companies or vendors should be involved in the testing of various services or, at a minimum, should review and certify any testing data on which the FCC intends to rely. The sample size should be large enough to be statistically significant and the protocol should measure accuracy for a set of minority groups who are at risk of being underserved.
- A baseline level of performance should be established that reflects the current service quality levels achieved by conventional IP CTS providers. ASR-only services must provide accuracy that is at a level "comparable" to that of the conventional services.
- Sample "development test" data could be made available to potential providers that would be representative of (but not identical to) the speech data to be used in the assessment evaluation. Providers would be allowed to train their systems on any data that they wish to use and that they can obtain legally.
- Testing of new ASR products should be conducted on a regular basis (perhaps once or twice per year). The evaluation data should not be seen by providers and potential providers. Evaluation data from a given year will be released to the public and may be used as developmental data in future evaluations.

- The group should exist on a continuing basis, and should continue to refine the metrics, datasets, methodologies, and standards—the latter of which should become more difficult as time progresses and technology progresses. The Commission could consider requiring providers to seek recertification on a periodic basis (for example, every 2 years) to show that they can satisfy the then-existing service quality standards.

V. SUMMARY AND CONCLUSIONS

In summary, I believe that absent further testing and technological innovation, existing ASR technology is not yet ready for incorporation into potential ASR-only IP CTS networks. Although the current systems attain remarkably good performance in baseline conditions, recognition accuracy for multiple populations that produce variant speech can be much worse. In addition, the systems are fragile to changes in the acoustical environment, and there is not yet a viable solution to these varied, complex, and likely compounding problems. I expect that as technology continues to develop, there will be a role for ASR-only IP CTS in the future. I conclude with a set of suggestions concerning how to establish an evaluation and certification process for new IP CTS providers that will facilitate innovation while also protecting users from sub-standard or discriminatory services.

BIBLIOGRAPHY

[1] X. Huang, A. Acero and H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Pearson Press, 2001.

[2] L. R. Rabiner and R. W. Schafer, Theory and Applications of Digital Speech Processing, Pearson, 2010.

[3] L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, Englewood Cliffs, NJ: Pearson, 1993.

[4] H. Hermansky, D. P. W. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000.

[5] Y. Miao and F. Metze, "End-to-end architectures for speech recognition," in *New Era for Robust Speech Recognition*, Cham, Springer Nature, 2017.

[6] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, Cambridge, MA: MIT Press, 2016.

[7] M. Nielsen, "Neural Networks and Deep Learning," 2016. [Online]. Available: http://neuralnetworksanddeeplearning.com.

[8] S. Goldwater, D. Jurafsky and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication,* vol. 52, no. 3, pp. 181-200, 2010.

[9] R. Tatman, "Gender and Dialect Bias in YouTube’s Automatic Captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Valencia, Spain, 2017.

[10] J. Hirschberg, D. Litman and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43, no. 1, pp. 155-175, 2004.

[11] A. Potamianos, S. Narayanan and S. Lee, "Automatic speech recognition for children," in *Eurospeech 1997*, Rhodes, Greece, 1997.

[12] J. G. Wilpon and C. N. Jacobsen, "A study fof speech recognition for children and the elderly," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, 1996.

[13] R. Vipperla, "Automatic speech recognition for ageing voices," Ph.D. Thesis School of Informatics, University of Edinburgh, Edinburgh, Scotland, 2011.

[14] M. Benzeghiba, R. De Mori,). Deroo, S. Dupont, T. Erbes, D. Jouvey, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyahi and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, pp. 763-786, 2007.

[15] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," *J. Acousti. Soc. Amer.*, vol. 116, no. 6, pp. 3647-3658, 2004.

[16] M. Sumner and A. G. Samuel, "The effect of experience on the perception and represen- tation of dialect variants.," *Journal of Memory and Language*, vol. 60, no. 4, pp. 487-501, 2009.

[17] W. Knight, "AI programs are learning to exclude some African-American voices," *MIT Technology Review*, 16 August 2017.

[18] R. Tatman and C. Kasten, "Effect of talker dialect, gender & race on accuracy of Bing speech and YouTube automatic captions," in *Interspeech 2017*, Stockholm, Sweeden, 2017.

[19] T. T. Peng, "Automatic Speech Recognition for Non-Native Speakers," Ph.D. thesis, Université Joseph-Fourier, Grenoble, France, 2008.

[20] J. K. A. S. Zapata, "Assessing the performance of automatic speech recognition systems when used by native and non-native speakers of three major languages in dictation workflows," in *Proc. 20th Nordic Conference of Computational Linguistics (NoDaLiDa) 2015)*, Vilnius, Lithuania, 2015.

[21] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429-442, 2000.

[22] A. Glasser, K. Kushalnagar and R. Kushalnagar, "Feasibility of using automatic speech recognition with voices of deaf and hard-of-hearing individuals," in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 17)*, Baltimore, Maryland, 2017.

[23] L. Moro-Velazquez, J. Cho, W. S., M. A. Hasegawa-Johnson, O. Scharenborg, H. Kim and N. Dehak, "Study of the performance of automatic speech recognition systems in speakers with Parkinson's Disease," in *Interspeech 2019*, Graz, Austria, 2019.

[24] D. Harwell, "Why some accents don't work on Alexa or Google Home," 19 July 2018. [Online]. Available: http://www.washingtonpost.com. [Accessed 19 September 2019].

[25] Johnson, "In the world of voice recognition, not all accents are equal," 15 February 2018. [Online]. Available: https://www.economist.com/books-and-arts/2018/02/15/in-the-world-of-voice-recognition-not-all-accents-are-equal. [Accessed 19 September 2019].

[26] S. Paul, "Voice is the next big platform unless you have an accent," March 2017. [Online]. Available: https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/. [Accessed 19 September 2019].

[27] B. Raj, V. N. Parikh and R. M. Stern, "The effects of background music on speech recognition accuracy," in *International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997.

[28] R. M. Stern, C. Kim, A. R. Moghimi and A. Menon, "Binaural technology and automatic speech recognition," in *International Conference on Acoustics*, Buenos Aires, Argentina, 2016.

[29] B. H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language,* vol. 5, pp. 275-294, 1991.

[30] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication,* vol. 22, pp. 1-15, 1997.

[31] D. Wang, M. Li and Z. Di, "True reason for Zipf's law in language," *Physica A: Statistical Mechanics and its Applications,* vol. 358, no. 2-4, pp. 545-550, 2005.

[32] K. C. Sim, Y. Qian, G. Mantena, L. Samarakoon, S. Kundu and T. Tan, "Adaptation of deep neural network acoustic models for robust automatic speech recognition," in *New Era for Robust Speech Recognition*, Cham, Springer, 2017.

[33] J. R. Pierce, "Whither Speech Recognition," *J. Acoust. Soc. Amer.*, vol. 46, no. 4, pp. 1049-1051, 1969.

[34] M. Y. Liberman, "Clinical applications of human language technology: opportunities and challenges," in *MSP-Podcast Workshop, University of Texas at Dallas*, Dallas, 2019.

APPENDIX. BRIEF SUMMARY OF RICHARD STERN'S CAREER

A curriculum vitae that summarizes my education, work history and publications is attached to this report as Exhibit 1.

I have been a professor at Carnegie Mellon University since January 1977, and have been working in the area of automatic speech recognition and related technologies since 1982. Specifically, I have studied and performed research in the fields of automatic speech recognition, auditory perception, general signal processing, and biomedical instrumentation.

I received a Bachelor of Science in Electrical Engineering from the Massachusetts Institute of Technology (MIT) in Cambridge, MA, in 1970. I received a Master of Science in Electrical Engineering and Computer Sciences from the University of California at Berkeley in 1972. In 1977, I was awarded a Ph.D. in Electrical Engineering and Computer Science, also by MIT. I began work immediately after as an Assistant Professor of Electrical Engineering at CMU.

I am currently a Full Professor in CMU's Department of Electrical and Computer Engineering. For a number of years I have also served as Professor by Courtesy in CMU's Language Technologies Institute, and the Department of Computer Science and Biomedical Engineering Departments, and I currently retain these positions. Since 2008 I have also served as a Lecturer in the CMU School of Music, where I manage the School's Programs in Music and Technology in collaboration with other faculty members.

Since 1982, my research has been focused on automatic speech recognition (ASR) technologies, and specifically the development of algorithms by which ASR systems can be made more robust in difficult acoustical environments.

I am one of the small number of individuals who has been elevated to the rank of Fellow by the Institute of Electrical and Electronics Engineers (IEEE), the International Speech

Communication Association (ISCA), and the Acoustical Society of America (ASA), and I have served in various positions of leadership in these organizations for some years. Among other positions and honors, I was the ISCA Distinguished Lecturer in 2008-2009. I also organized and served as General Chair for Interspeech 2006, the largest professional conference for researchers in speech and natural language science and technologies.

Between approximately 1990 and 1995 I served in various positions of leadership for the Defense Advanced Research Projects Agency (DARPA) Program in Speech and Natural Language Technologies. During this time I served as Secretary of the DARPA Spoken Language Coordinating Committee, chaired the subcommittee that established the standards used to define the performance of speech recognition systems, and organized and chaired a number of the DARPA conferences at which progress was reported and new evaluation metrics were presented and discussed. Through these experiences I obtained extensive first-hand knowledge of the development of objective evaluation standards for speech recognition systems, as well as the impact of the community-wide acceptance and adoption of these standards on progress in the field.

I continue to remain active in the development of new speech processing technologies to this day.

EXHIBIT 1

CURRICULUM VITA

Richard M. Stern, Jr.

Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
Phone: (412) 268-2535
FAX: (412) 268-3890

513 Emerson Street
Pittsburgh, PA 15206
Cell phone: (412) 916-7386

Email: rms@cs.cmu.edu
Citizenship: U.S.A.

www.ece.cmu.edu/~rms

PROFESSIONAL INTERESTS Automatic speech recognition, auditory perception, acoustics, signal processing, biomedical instrumentation

EDUCATION

Ph.D. (1977)	Electrical Engineering and Computer Science Massachusetts Institute of Technology, Cambridge, MA
M.S. (1972)	Electrical Engineering and Computer Sciences University of California, Berkeley, CA
S.B. (1970)	Electrical Engineering Massachusetts Institute of Technology, Cambridge, MA

EXPERIENCE

1995 - present	Professor of Electrical and Computer Engineering Carnegie Mellon University.
1988 - present	Associate Professor and Professor by Courtesy, Language Technologies Institute, Computer Science Department, Biomedical Engineering Department
2009 - present	Lecturer, School of Music Carnegie Mellon University
1995 - 2003	Associate Director of the Information Networking Institute Carnegie Mellon University
1982 - 1995	Associate Professor of Electrical and Biomedical Engineering Carnegie Mellon University
1985	Visiting Professor in Speech and Communication Sciences, Nippon Telegraph and Telephone Electrical Communications Laboratory, Tokyo, Japan
1977 - 1982	Assistant Professor of Electrical and Biomedical Engineering Carnegie Mellon University

1979 - 1981	Adjunct Assistant Professor of Otolaryngology University of Pittsburgh School of Medicine
1973 - 1976	Teaching and Research Assistant, Department of Electrical Engineering, Massachusetts Institute of Technology

PROFESSIONAL ACTIVITIES (partial listing)

Distinguished Lecturer, International Speech Communication Association, 2008-2009.

General Chair, INTERSPEECH International Conference on Spoken Language Processing, September, 2006.

Technical Program Co-Chair, IEEE Workshop on Automatic Speech Recognition and Understanding, December 2005.

Technical Program Chair, 141st meeting of the Acoustical Society of America, June 2002.

General Chair, DARPA Spoken Language Technologies Workshop, March, 1994.

Publications Chair, ARPA Spoken Language Technology and Applications Day, April, 1993.

Publications Chair, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October, 1993.

Chair, standing DARPA Speech and Natural Language Workshop Organizing Committee, 1991 - 1992.

Secretary, ARPA Spoken Language Coordinating Committee, 1990 - 1995.

General Chair, DARPA Speech and Natural Language Workshop, June, 1990.

International Advisory Board, International Speech Communication Association, 2006 - present.

International Advisory Board, Center for Speech and Language Technologies, Tsinghua University, Beijing, China, 2007 - 2010.

Chair, Selection Committee for IEEE James L. Flanagan Speech & Audio Processing Award, 2006 - 2008.

IEEE Signal Processing Society Technical Committee on Audio and Electroacoustics, 1991 - 1995.

IEEE Signal Processing Society Technical Committee on Speech, 1993 - 1997.

Editorial board, *Journal of Computer Speech and Language*, 1994 - present.

Editorial board, *Free Speech Journal*, 1996 - 1998.

Ongoing collaborative research in binaural hearing with the Department of Otolaryngology at the University of Connecticut Medical School, Farmington, CT.

Member of Institute of Electrical and Electronics Engineers, Acoustical Society of America,

International Speech Communication Association, Association for Research in Otolaryngology, Audio Engineering Society

Reviewer for National Science Foundation, International Speech Communication Association, IEEE, *J. Acoust. Soc. Amer.*, *Hearing Research*, *IEEE Transactions on Signal Processing*, *IEEE Transactions on Speech and Language*, *IEEE Transactions on Systems, Man, and Cybernetics*, and *Communications of the Association of Computing Machinery.*

HONORS AND AWARDS

Fellow, Institute of Electrical and Electronics Engineers (IEEE)

Fellow, International Speech Communication Association (ISCA)

Fellow, Acoustical Society of America (ASA)

Distinguished Lecturer of the International Speech Communication Association, 2008 to 2009

Allen Newell Award for Research Excellence, Carnegie Mellon University Department of Computer Science, 1992

Lutron Award for Teaching Excellence, Carnegie Mellon University Department of Electrical and Computer Engineering, 2018

IEEE Student Branch Award for Teacher of the Year, Carnegie Mellon University Department of Electrical Engineering, 1979

PUBLICATIONS AND PAPERS

Papers in Archival Journals

STERN, R. M., COLBURN, H. S., BERNSTEIN, L. R., AND TRAHIOTIS, C. (2019). "The fMRI data of Thompson *et al.* (2006) do not constrain how the human midbrain represents interaural time delay," *Journal of the Association for Research in Otolaryngology* **20**:305-311.

DIETZ, M., LESTANG, J.-H., MAJDAK, P., STERN, R. M., MARQUARDT, T., EWERT, S. D., HARTMANN, W. M., and GOODMAN, D. F. M. (2017). "A Framework for Testing and Comparing Binaural Models," *Hearing Research* **360**:92-106.

DE LA CALLE SILOS, F., and STERN, R. M. (2017). "Synchrony-based feature extraction for robust automatic speech recognition," *IEEE Signal Processing Letters* **24:**1158-1162.

FREDES, J., NOVOA, J., KING, S., STERN, R. M., and BECERRA YOMA, N. (2017). "Locally-normalized filter banks applied to deep neural network-based robust speech recognition," *IEEE Signal Processing Letters* **24:**377-381.

KIM, C., and STERN, R. M. (2016). "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing* **24:**1315-1329.

CHO, B. J., KWON,H. Cho, J.-W., KIM, C., STERN, R. M. and PARK, H.-M. (2016). "A subband-based stationary-component suppression method using harmonics and power ratio for reverberant speech recognition, *IEEE Signal Processing Letters*, **23**:780-784.

ROMIGH, G. D., BRUNGART, D. S., STERN, R. M., and SIMPSON, B. D. (2015). "Efficient real

spherical harmonic representation of head-related transfer functions," *IEEE Journal of Selected Topics in Signal Processing*, **9**: 921-930, August 2015.

POBLETE, V., ESPIC, F., KING, S., STERN, R. M., HUENEPAN, F., and BECERRA YOMA, N. (2015). "A perceptually-motivated low-complexity channel normalization technique applied to speaker verification." *Computer Speech and Language,* **31**1-27, 2015.

POBLETE, V., BECERRA YOMA, N., and STERN, R. M. (2014). "Optimizing the parameters characterizing sigmoidal rate-level functions based on acoustic features," *Speech Communication,* **56:**19-34, January 2014.

HERMANSKY, H., COHEN, J. R., and STERN, R. M. (2013). "Perceptual properties of current speech recognition technology," *Proc. IEEE* **101**:1968-1985, September 2013.

STERN, R. M., and MORGAN, N. (2012). "Hearing is believing: biologically-inspred methods for robust speech recognition," *IEEE Signal Processing Magazine* **29**:34-43, November, 2012.

CHIU, Y.-H. B., RAJ, B., and STERN, R. M. (2012). "Learning-based auditory encoding for robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing* **20:**900-914, March 2012.

KIM, W., and STERN, R. M. (2011). "Mask classification for missing-feature reconstruction for robust speech recognition," *Speech Communication*, **53:**1-11, January 2011.

PARK, H.-M., and STERN, R. M. (2009). "Spatial Separation of Speech Signals using Continuosly-Variable Weighting Factors Estimated from Comparisons of Zero Crossings, *Speech Communication Journal,* **51**(1):15-25, January 2009.

SELTZER, M. L., and STERN, R. M. (2006). "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments," *IEEE Transactions of Speech, Language, and Audio Processing* 14(6): 2109-2121, November 2006.

RAJ, B., and STERN, R. M. (2005). "Missing-Feature Methods for Robust Automatic Speech Recognition," *IEEE Signal Processing Magazine*, September 2005.

KIM, N. S., LIM, W., and STERN, R. M. (2005). "Feature compensation based on switching linear dynamic model," *IEEE Signal Processing Letters*, **12**(6): 473-476.

SELTZER, M. L., RAJ, B., and STERN, R. M. (2004). "Likelihood-Maximizing Beamforming for Robust Hands-Free Speech Recognition," *IEEE Transactions of Speech and Audio Processing*, **12**(5): 489-498, September 2004.

OBUCHI, Y., HATAOKA, N., and STERN, R. M. (2004), "Normalization of Time-Derivative Parameters for Robust Speech Recognition in Small Devices," *IEICE Trans. on Information and Systems*, **87-D**(4): 1004:1011, April 2004.

RAJ, B., SELTZER, M. L., and STERN, R. M. (2004), "Reconstruction of Missing Features for Robust Speech Recognition," *Speech Communication Journal*, **43**(4): 275-296, September 2004.

SELTZER, M. L., RAJ, B,, and STERN, R. M. (2004). "A Bayesian Framework for Spectrographic Mask Estimation for Missing Feature Speech Recognition," *Speech Communication Journal,* **43**(4): 379-393, September 2004.

SINGH, R., RAJ, B., and STERN, R. M. (2001), "Automatic Generation of Sub-Word Units for

Speech Recognition Systems," *IEEE Trans. on Speech and Audio Proc.* **10**(2):89-99.

HUERTA, J. M., and STERN, R. M. (2001). "Distortion-Class Modeling for Robust Speech Recognition under GSM RPE-LTP Coding," *Speech Communication Journal*, **34:**213-225 (invited paper).

MORENO, P. J., RAJ, B., and STERN, R. M. (1998). "Data-Driven Environmental Compensation for Speech Recognition: A Unified Approach," *Speech Communication Journal*, **24:** 267-85.

STERN, R. M., and SHEAR, G. D. (1996a) "Lateralization and Detection of Low- Frequency Binaural Stimuli: Effects of Distribution of Internal Delay," *J. Acoust. Soc. Amer.* **100**: 2278-2288.

STERN, R. M., and SHEAR, G. D. (1996b) "Lateralization and Detection of Low- Frequency Binaural Stimuli: Specification of the Extended Position-Variable Model," *Physics Auxiliary Publication Service*, AIP document E-JASMA-100-2278- 0.175MB via http://www.aip.org/epaps/epaps.html.

TRAHIOTIS, C., and STERN, R. M. (1994) "Across-Frequency Interaction in Lateralization of Complex Binaural Stimuli," *J. Acoust. Soc. Amer.* **96**: 3804- 3806 (L).

STERN, R. M., ZEPPENFELD, T., and SHEAR, G. D. (1991). "Lateralization of Rectangularly-Modulated Noise: An Explanation for Counterintuitive Reversals," *J. Acoust. Soc. Amer.* **90:** 1901-1907.

COAST, D. A., STERN, R. M., CANO, G. G., and BRILLER, S. A. (1990). "An Approach to Cardiac Arrhythmia Analysis Using Hidden Markov Models," *IEEE Trans. Biomed. Eng*. **37**: 826-836.

TRAHIOTIS, C., and STERN, R. M. (1989). "Lateralization of Bands of Noise: Effects of Bandwidth and Differences of Interaural Time and Phase," *J. Acoust. Soc. Amer.* **86**: 1285-1293.

RUDNICKY, A. I., and STERN, R.M. (1989). "Spoken Language Research at Carnegie Mellon," *Speech Technology Magazine* **4:** 38-43.

STERN, R. M., ZEIBERG, A. S., and TRAHIOTIS, C. (1988). "Lateralization of Complex Binaural Stimuli: A Weighted Image Model," *J. Acoust. Soc. Amer.* 84, 156-165.

STERN, R. M., and LASRY, M. J. (1987). "Dynamic Speaker Adaptation for Feature-Based Isolated Letter Recognition," I*EEE Trans. on Acoustics, Speech, and Signal Processing* **35:** 751-763.

STERN, R. M., and COLBURN, H. S. (1985). "Lateral-Position Models of Interaural Discrimination," J*. Acoust. Soc. Amer.* **77:** 753-755.

STERN, R. M., and COLBURN, H. S. (1985). "Subjective Lateral Position and Interaural Discrimination," *Physics Auxiliary Publication Service*, AIP document no. PAPS JASMA-77-753-29.

LASRY, M. J., and STERN, R. M. (1984). "A Posteriori Estimation of Correlated Jointly Gaussian Mean Vectors," *IEEE Trans. on Pattern Anal. and Mach. Intel.* **6:** 530-535.

CROWLEY, J. L., and STERN, R. M., Jr. (1984). "Fast Computation of the Difference of Low Pass (DOLP) Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6:** 212-222.

STERN, R. M., Jr., SLOCUM, J. E., and PHILLIPS, M. S. (1983). "Interaural Time and Amplitude Discrimination in Noise", *J. Acoust. Soc. Amer.* **73:**1714-1722.

YOST, W. A., GRANTHAM, D. W., LUFTI, R. A., and STERN, R. M., Jr. (1982). "The Phase Angle of Addition in Temporal Masking for Diotic and Dichotic Listening Conditions," *Hearing Res.* **7:** 247-259.

MURTI, K. G., STERN, R. M., CANTEKIN, E. I. and BLUESTONE, C. D. (1982). "Classification of Spectral Patterns Obtained from Eustacian Tube Sonometry," *IEEE Trans. Biomed. Eng.* **29:** 473-477.

MURTI, K. G., STERN, R. M., Jr., CANTEKIN, E. I. and BLUESTONE, C. D. (1980). "Sonometric Evaluation of Eustachian Tube Function Using Broadband Stimuli", *Annals of Otology, Rhinology, and Laryngology*, (Suppl. 68) 89, 178-189.

RUOTOLO, B. R., STERN, R. M., Jr., and COLBURN, H. S. (1979). "Discrimination of Symmetric, Time- Intensity Traded Binaural Stimuli," *J. Acoust. Soc. Amer.*, **66:** 1733-1737.

STERN, R. M., Jr. and COLBURN, H. S. (1978). "Theory of Binaural Interaction Based on Auditory-Nerve Data. IV. A Model for Subjective Lateral Position," *J. Acoust. Soc. Amer.*, **64:** 127-140.

Critically-Reviewed Books, Book Chapters, and Theses

STERN, R. M. and MENON, A. (2019). "Binaural Technology for Machine Speech Recognition and Understanding," Chapter in *The Technology of Binaural Understanding*, J. Blauert and J. Braasch, Eds., Springer Nature.

MITRA, V., FRANCO, H., STERN, R. M., VAN HOUT, J., FERRER, L.,GRACIARENA, M., WANG, W., VERGYRI, D., ALWAN, A., and HANSEN, J.H.L. (2017). "Robust features in Deep Learning based Speech Recognition," Chapter in *New Era for Robust Speech Recognition: Exploiting Deep Learning*, S. Watanabe, M. Delcroix, F.Metze, & J. R. Hershey (eds.), pp. 165-196, Springer International Publishing.

STERN, R. M. and MORGAN, N. (2013). "Features Based on Auditory Physiology and Perception," Chapter in *Noise-Robust Techniques for Automatic Speech Recognition,* T. Virtanen, R. Singh, and B. Raj, Eds., Wiley Press.

STERN, R. M., WANG, D., and BROWN, G. (2006). "Binaural Sound Localization," Chapter in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. Wang and G. Brown, Eds., Wiley and IEEE Press.

STERN, R. M., TRAHIOTIS, C., and RIPEPI, A. M. (2006). "Fluctuations in Amplitude and Frequency Enable Interaural Delays to Foster the Identification of Speech-like Stimuli," Chapter in *Dynamics of Speech Production and Perception*, P. Divenyi, Ed., IOS Press.

TRAHIOTIS, C., BERNSTEIN, L. R., STERN, R. M., and BUELL, T. N. (2005). "Interaural Correlation as the Basis of a Working Model of Binaural Processing: An Introduction," Chapter in *Springer Handbook of Auditory Research: Sound Source Localization*, R. Fay and T. Popper, Eds., Springer-Verlag.

STERN, R. M. (2004). "Signal Separation Motivated by Human Auditory Perception: Applications to Automatic Speech Recognition," Chapter in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., Springer-Verlag.

SINGH, R., STERN, R. M., and RAJ, B. (2002). "Signal and Feature Compensation Methods for Robust Speech Recognition," Chapter in *CRC Handbook on Noise Reduction in Speech Applications*, Gillian Davis, Ed., Boca Raton: CRC Press.

SINGH, R., RAJ, B., and STERN, R. M. (2002). "Model Compensation and Matched Condition Methods for Robust Speech Recognition," Chapter in *CRC Handbook on Noise Reduction in Speech Applications*, Gillian Davis, Ed., Boca Raton: CRC Press.

STERN, R. M., ACERO, A., LIU, F.-H., and OHSHIMA, Y. (1996). "Signal Processing for Robust Speech Recognition," Invited chapter in *Speech Recognition*, pp. 351-378, C.-H. Lee and F. Soong, Eds., Boston: Kluwer Academic Publishers.

STERN, R. M., and TRAHIOTIS, C. (1996). "Models of Binaural Perception," Invited chapter in *Binaural and Spatial Hearing in Real and Virtual Environments*, pp. 499-531, R. Gilkey and T. R. Anderson, Eds. New York: Lawrence Erlbaum Associates

STERN, R. M. (1995). "Robust Speech Recognition," Invited chapter in *Survey on the State of the Art in Speech and Natural Language Processing*, R. A. Cole et al., Ed.

STERN, R. M., and TRAHIOTIS, C. (1995). "Models of Binaural Interaction," Invited chapter in *Handbook of Perception and Cognition, Volume 6: Hearing*, pp. 347-386, B. C. J. Moore., Ed. New York: Academic Press.

STERN, R. M., Jr. (1976b). *Lateralization, Discrimination, and Detection of Binaural Pure Tones*, Ph.D. Thesis, Electrical Engineering Department, MIT, December, 1976.

Invited Conference Presentations

STERN, R. M. (2017). "Predicting Binaural Lateralization, Interaural Discrimination, and Binaural Detection Using the Position-Variable Model," invited talk at the Macquarie University ARC Laureate Workshop: Creating a Sense of Auditory Space, October 2017, Sydney, Australia.

STERN, R. M, KIM, C., MOGHIMI, A.R., and MENON, A. (2016). "Binaural technology and automatic speech recognition," invited talk at the International Congress on Acoustics, September 2016, Buenos Aires, Argentina.

STERN, R. M. (2016). "Applying Models of Auditory Processing to Automatic Speech Recognition: Progress and Promise," invited keynote talk at the 2015 Meeting of the Information Processing Sociaty of Japan: Special Interest Group in Spoken Language Processing, Toyama, Japan, February, 2016.

STERN, R. M. (2014). "Applying Models of Auditory Processing to Automatic Speech Recognition: Progress and Promise," invited talk at the Frederick Jelinek Memorial Workshop on Meaning Representations in Language and Speech Processing, Prague, Czech Republic, July, 2014.

STERN, R. M. (2014). "Robust Automatic Speech Recognition in the 21st Century," invited keynote talk at the 2014 AFEKA Conference for Speech Processing, Tel Aviv, Israel, July, 2014.

STERN, R. M. (2014). "Applying Models of Auditory Processing to Automatic Speech Recognition: Promise and Progress," International Symposium on Speech Recognition, University of Zaragoza, Spain, May, 2014.

STERN, R. M. (2011). "Applying Physiologically-Motivated Models of Auditory Processing to Automatic Speech Recognition," invited talk at the Third International Symposium on Auditory

and Audiological Research, Nyborg, Denmark, August, 2011.

STERN, R. M. (2010). "The impact of the distribution of internal delays in binaural models on predictions for psychoacoustical data," invited talk at the 161th Meeting of the Acoustical Society of America, Cancun, Mexico, November, 2010.

STERN, R. M. (2009). "New Directions in Robust Speech Recognition: What We Can Learn from Auditory Models," invited keynote address at the Symposium on Frontiers of Research in Speech and Music, Gwalior, India, December, 2009.

STERN, R. M. (2009). "New Directions in Robust Automatic Speech Recognition," invited keynote address at the Workshop on Image and Speech Processing, Hyderabad, India, December, 2009.

STERN, R. M. (2008). "Applying Physiologically-Motivated Models of Auditory Processing to Automatic Speech Recognition: Promises, Progress, and Problems," Invited keynote address at the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, Brisbane, Australia, September, 2008.

STERN, R. M, GOUVEA, E., KIM, C., KUMAR, K., and PARK, H.-M. (2008). "Binaural and Multiple-Microphone Processing for Robust Automatic Speech Recognition," Invited keynote address at the IEEE Workshop on Hands-free Speech Communication and Microphone Arrays, Trento, Italy, May, 2008.

STERN, R. M. (2004). "Signal Processing for Sound Separation and Robust Representation," Invited keynote address at AFOSR/NSF Symposium on Speech Separation and Comprehension in Complex Acoustic Environments, Montreal, Quebec, November 2004.

STERN, R. M. (2003). "Signal Separation Motivated by Auditory Processing: Applications to Speech Recognition," invited review talk at the NSF Symposium on Signal Separation, Montreal, Quebec, November, 2003.

STERN, R. M. (2003). "Signal Processing for Robust Recognition," invited talk at the NAIST International Center of Excellence Symposium, Nara, Japan, March, 2003.

STERN, R. M. (2002). "Using Computational Models of Binaural Hearing to Improve Automatic Speech Recognition Accuracy: Promise, Progress, and Problems," AFOSR Workshop on Computational Audition, Columbus, Ohio, August, 2002.

STERN, R. M. (2000). "Robust Signal Representations for Automatic Speech Recognition," Institute for Mathematics and Its Applications Workshop on the Mathematical Foundations of Speech Processing and Recognition, Minneapolis, Minnesota, September, 2000.

STERN, R. M. (2000). "The Language of Music," invited keynote talk presented at the Third International Symposium on Text, Speech, and Dialog, Brno, Czech Republic, September, 2000.

STERN, R. M. (2000). "Tendencias Actuales en el Procesamiento del Lenguaje Hablado y Sistemas Conversacionales (Current Trends in Spoken Language Processing and Conversational Systems)", invited keynote talk at the XV Simposium Internacional de Electrónica y Comunicación, Instituto Tecnológico de Estudios Superiores de Monterrey Mexico, February, 2000.

STERN, R. M. (1999). "Tendencias Actuales en el Procesamiento del Lenguaje Hablado y Sistemas Conversacionales (Current Trends in Spoken Language Processing and Conversational Systems)", invited keynote talk at the XXIV Simposium Internacional de Sistemas Computacio-

nales, Instituto Tecnológico de Estudios Superiores de Monterrey, Monterrey, Mexico, March, 1999.

STERN, R. M., and TRAHIOTIS, C. (1997). "Binaural Mechanisms that Emphasize Consistent Interaural Timing Information over Frequency," invited keynote talk in *Psychophysical and Physiological Advances in Hearing,* Proceedings of the XI International Symposium on Hearing, August, 1997, Grantham, United Kingdom. A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis, Eds., Whurr Publishers, London, 1998.

STERN, R. M., RAJ, B., and MORENO, P. J. (1997). "Compensation for Environmental Degradation in Automatic Speech Recognition," invited keynote talk presented at the *Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, April, 1997, Pont-au-Mousson, France, pp. 33-42.

STERN, R. M. (1996). "The Current State of the Art of in Speech Recognition (Estado-da-Arte em Reconhecimento de Voz)," invited keynote talk presented at VOICETECH'96, the First Brazilian Workshop in Automatic Speech Recognition Campinas, Sao Paolo, Brazil, September, 1996.

STERN, R. M. (1996). "New Directions in Spoken Language Processing," invited talk at the Second Joint NSF/CONACyT Workshop on Bilateral Collaboration, Jalapa, Mexico, March, 1996.

STERN, R. M. (1996). "Tendencias Actuales en el Procesamiento del Lenguaje Hablado (Current Trends in Spoken Language Processing)", invited talk at the Universidad Veracruzana, Jalapa, Mexico, March, 1996.

STERN, R. M., and SULLIVAN, T. M. (1996). "Robust Speech Recognition Using Signal Processing Based On Binaural Perception," invited talk presented at the First Forum Acusticum, Antwerp, Belgium, April, 1996.

STERN, R. M., MORENO, P.J., and RAJ, B. (1996). "Compensation for Speech Recognition in Degraded Acoustical Environments," invited talk at the 132th meeting of the Acoustical Society of America, Honolulu, Hawaii, December, 1996.

STERN, R. M. (1995). "Nuevos Enfoques en Procesamiento de Lenguaje Hablado (New Directions in Spoken Language Processing)," invited talk at the Universitat Politecnica de Catalunya, Barcelona, Spain, September, 1995.

STERN, R. M. (1995). "New Directions in Spoken Language Processing," invited talk presented at the Telefónica Investigación y Desarrollo Laboratory Symposium on Spoken Language Processing, Madrid, Spain, September 1995.

STERN, R. M. (1995). "Automatic Speech Recognition using Signal Processing based on Auditory Physiology and Perception," invited paper presented at the 129th meeting of the Acoustical Society of America, Washington, D.C., June, 1995.

MORENO, P. J., RAJ, B., and STERN, R. M. (1995). "Approaches to Environmment Compensation in Automatic Speech Recognition," invited paper presented at the 15th International Conference on Acoustics, Trondheim, Norway, Vol. III, pp. 109-112, June, 1995.

STERN, R. M., and SULLIVAN, T. M. (1994). "Robust Speech Recognition Based on Human Binaural Perception," invited paper presented at the ATR workshop on A Biological Framework for Speech Perception and Production, Kansai Science City, September, 1994. Reprinted in

ATR technical report TR-H-121: *Proceedings of the ATR workshop on A Biological Framework for Speech Perception and Production*, 122 pages, (1995).

STERN, R. M. LIU, F.-H., SULLIVAN, T. M., MORENO, P. J., and ACERO, A. (1994). "Multiple Approaches to Robust Speech Recognition," invited keynote paper at the Fifth Western Pacific Regional Acoustical Conference, Seoul, Korea, August, 1994.

STERN, R. M. (1993). "Models of Binaural Interaction," invited keynote paper at the AFOSR Conference on Binaural and Spatial Hearing, Wright-Patterson Air Force Base, September, 1993.

STERN, R. M. (1993). "Psychoacoustical Basis of Machine Speech Recognition," invited talk at the Annual Meeting of the American Association for the Advancement of Science, February, 1993.

STERN, R. M. (1989). "Recent Progress in Spoken-Language Systems," invited lecture at the Second International Symposium on Artificial Intelligence, Monterrey, Mexico, October, 1989.

STERN, R. M. (1988). "Overview of Models of Binaural Perception," invited review paper at the 1988 National Research Council CHABA Symposium, Washington, D.C., October, 1988.

STERN, R. M. (1988). "Estado Actual de la Tecnología de Entradas/Salidas de Canales de Voz (Overview of Current Voice Input/Output Technologies)," invited keynote lecture at the XIII Simposium Internacional de Sistemas Computacionales, Monterrey, Mexico, March, 1988.

COLE, R. A., STERN, R. M., and LASRY, M. J. (1986). "Performing Fine Phonetic Distinctions: Templates vs. Features," invited talk, reprinted in *Invariance and Variability of Features in Spoken English Letters*, J. Perkell *et al*., eds., Lawrence Erlbaum, New York.

Critically-Reviewed Conference Presentations

MENON, A., Kim, C., and STERN, R. M. (2019), "Robust recognition of reverberant and noisy speech using coherence-based processing," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2019, Brighton, United Kingdom.

XIA, Y. and STERN, R. M. (2018). *A Priori* SNR Estimation Based on a Recurrent Neural Network for Robust Speech Enhancement, *Interspeech 2018*, September 2018, Hyderabad, India.

MICHELSON, J., STERN, R. M., and SULLIVAN, T. M. (2018), "Automatic guitar tablature transcription from audio using inharmonicity regression and bayesian classification," *145th COnvention of the Audio Engineering Society*, October 2018, New York City.

KIM, C., MENON, A., BACCHIANI, M., and STERN, R. M. (2018), "Source separation using phase difference and reliable mask selection," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2018, Calgary, Alberta, Canada.

MENON, A., KIM, C., KUROKAWA, U., and STERN, R. M. (2017), "Binaural Processing for Robust Recognition of Degraded Speech," *IEEE Automatic Speech Recognition and Understanding Workshop,* December 2017, Naha, Okinawa, Japan.

MENON, A., KIM, C., and STERN, R. M. "Robust Speech Recognition Based on Binaural Auditory Processing*," Interspeech 2017*, September 2017, Stockholm, Sweden.

NOVOA, J., WUTH, J., ESCUDERO, J. P., FREDES, J., MAHU, R., STERN, R. M., and BECERRA YOMA, N., "Robustness over time-varying channels in DNN-HMM ASR-Based human-robot interaction," *Interspeech 2017,* September 2017, Stockholm, Sweden.

FREDES, J., NOVOA, J., POBLETE, V., KING, S., STERN, R. M., and YOMA, N. B. (2015), "Robustness to additive noise of locally-normalized cepstral coefficients in speaker verification," *Interspeech 2015*, September 2015, Dresden, Germany.

HARVILLA, M. H., and STERN, R. M. (2014). "Least squares declipping for robust speech recognition," *Interspeech 2014*, September 2014, Singapore.

MOGHIMI, A. R., RAJ, B., and STERN, R. M. (2014), "Post-masking: A hybrid approach to array processing for speech recognition," *Interspeech 2014*, September 2014, Singapore.

KIM, C., CHIN, K. K., BACCHIANI, M., and STERN, R. M. (2014), "Robust speech recognition using temporal masking and threshold algorithm," *Interspeech 2014*, September 2014, Singapore.

PARK, H.-M., MACIEJEWSKI, M., KIM, C., and STERN, R. M. (2014). "Robust speech recognition in reverberant environments using subband-based steady-state monaural and binaural suppression," *Interspeech 2014*, September 2014, Singapore.

MOGHIMI, A. R.,and STERN, R. M. (2014), "An analysis of binaural spectro-temporal masking as nonlinear beamforming," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2014, Florence, Italy.

HARVILLA, M., and STERN, R. M. (2012). "Histogram-based subband power warping and spectral averaging for robust speech recognition under matched and multistyle training," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2012, Kyoto, Japan.

KIM, C. and STERN, R. M. (2012). "Power-normalized cepstral coefficients (PNCC) for robust speech recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2012, Kyoto, Japan.

KIM, C., KHAWAND, C, and STERN, R. M. (2012). "Two-microphone source separation algorithm based on statistical modeling of angle distributions," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2012, Kyoto, Japan.

KIM, C., KUMAR, K., and STERN, R. M. (2011). "Binaural sound source separation motivated by auditory processing", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2011, Prague, Czech Republic.

KUMAR, K., KIM, C., and STERN, R. M. (2011). "Delta-spectral cepstral coefficients for robust speech recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2011, Prague, Czech Republic.

KUMAR, K., RAJ, B., SINGH, R., and STERN, R. M. (2011)."An iterative least-squares techique for dereverberation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2011, Prague, Czech Republic.

KUMAR, K., SINGH, R., RAJ, B., and STERN, R. M. (2011). "Gammatone sub-band magnitude-domain dereverberation, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2011, Prague, Czech Republic.

KIM, C., STERN, EOM, K., and Lee, J. (2010). “Automatic selection of thresholds for signal separation algorithms based on interaural delay,” *Interspeech 2010, September 2010,* Makuhari, Japan.

KIM, C., and STERN, R. M. (2010). “Nonlinear enhancement of onset for robust speech recognition,” *Interspeech 2010, September 2010,* Makuhari, Japan.

AL BAWAB, Z., RAJ, B, and STERN, R. M. (2010). “A hybrid physical and statistical dynamic articulatory framework incorporating analysis-by-synthesis for improved phone classification,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2010, Dallas, Texas.

CHIU, Y.-H., RAJ, B., and STERN, R. M. (2010). “Learning Based Auditory Encoding For Robust Speech Recognition,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2010, Dallas, Texas.

KIM, C., and STERN, R. M. (2010). “Feature Extraction For Robust Speech Recognition Based On Maximizing The Sharpness Of The Power Distribution And On Power Flooring,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2010, Dallas, Texas.

KUMAR, K., and STERN, R. M. (2010). “Maximum-Likelihood-Based Cepstral Inverse Filtering For Blind Speech Dereverberation,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2010, Dallas, Texas.

CHIU, Y.-H. B, and STERN, R. M. (2009). “Minimum variance modulation filters for robust speech recognition,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2009, Taipei, Taiwan.

AL BAWAB, Z., TURICCHIA, L., STERN, R. M., and RAJ, B. (2009). "Deriving vocal tract shapes from electromagnetic articulograph data via geometric adaptation and matching, *Interspeech 2009, September 2009,* Brighton, United Kingdom.

BUERA, L., MIGUEL, A., ORTEGA, E., LLEIDA, E., and STERN, R. (2009). "Unsupervised training scheme with non-stereo data for empirical feature vector compensation, *Interspeech 2009, September 2009,* Brighton, United Kingdom.

CHIU, Y.-H. B., RAJ, B., and STERN, R. M. (2009). "Learning-based auditory encoding," *Interspeech 2009, September 2009,* Brighton, United Kingdom.

GU, L., and STERN, R. M. (2009). "Speaker segmentation and clustering for sumultaneously-presented speech," *Interspeech 2009, September 2009,* Brighton, United Kingdom.

KIM, C., KUMAR, K., RAJ, B., and STERN, R. M. (2009). "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain," *Interspeech 2009, September 2009,* Brighton, United Kingdom.

KIM, C., and STERN, R. M. (2009). “Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction,” *Interspeech 2009, September 2009,* Brighton, United Kingdom.

KIM, C., and STERN, R. M. (2009). "Power Function-Based Power Distribution Normalization Algorithm for Robust Speech Recognition," *IEEE Automatic Speech Recognition and Understanding Workshop*, December 2009, Merano, Italy.

KIM, C., and STERN, R. M. (2009). "Robust Speech Recognition using a Small Power Boosting Algorithm," *IEEE Automatic Speech Recognition and Understanding Workshop*, December 2009, Merano, Italy.

CHIU, Y.-H., and STERN, R. M. (2008). "Analysis of Physiologically-Motivated Signal Processing for Robust Speech Recognition," Interspeech 2008, September 2008, Brisbane, Australia.

KIM, C., and STERN, A. M. (2008). "Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis," Interspeech 2008, September 2008, Brisbane, Australia.

AL BAWAB, Z., RAJ, B., and STERN, R. M. (2008). "Analysis-by-synthesis features for speech recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, April 2008, Las Vegas, Nevada.

GU, L., and STERN, R. M., "Single-channel speech separation based on modulation frequency," IEEE International Conference on Acoustics, Speech, and Signal Processing, April 2008, Las Vegas, Nevada.

KUMAR, K., and STERN, R. M. (2008). "Environment-invariant compensation for reverberation using linear post-filtering for minimum distortion," IEEE International Conference on Acoustics, Speech, and Signal Processing, April 2008, Las Vegas, Nevada.

STERN, R. M., GOUVEA, E., and THATTAI, G. (2007). "'Polyaural' array processing for automatic speech recognition in degraded environments," Interspeech 2007, August 2007, Antwerp, Belgium.

PARK, H.-M., and STERN, R. M. (2007). "Missing-feature speech recognition using dereverberation and echo supporession in reverberant environments," IEEE International Conference on Acoustics, Speech, and Signal Processing, April 2007, Honolulu, Hawaii.

KUMAR, K., CHEN, T., and STERN, R. M. (2007). "Profile view lip reading," IEEE International Conference on Acoustics, Speech, and Signal Processing, April 2007, Honolulu, Hawaii.

KIM, C., CHIU, Y.-H., and STERN, R. M. (2006). "Physiologically-motivated synchrony-based processing for robust automatic speech recognition," Interspeech 2006, September 2006, Pittsburgh, Pennsylvania.

NARAYANASWAMY, B., GANGAGHARIAH, R., and STERN, R. M. (2006). "Voting for two speaker segmentation," Interspeech 2006, September 2006, Pittsburgh, Pennsylvania.

PARK, H.-M., and STERN, R. M. (2006). "Spatial separation of speech sgnals using continuously-variable masks estimated from comparisons of zero crossings," IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2006, Toulouse, France.

KIM, W., and STERN, R. M. (2006). "Band-independent mask estimation for missing-feature reconstruction," IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2006, Toulouse, France.

KIM, W., STERN, R. M., and KO, H. (2005). "Environment-Independent Mask Estimation for Missing Feature Reconstruction," *Proc. Eurospeech-2005* September, 2005, Lisbon, Portugal.

LI, X., and STERN, R. M. (2004). "Parallel Feature Generation Based on Maximum Normalized Acoustic Likelihood for Improved Combination Performance," *Proc. of the International Conference of Spoken Language Processing*, October, 2004, Jeju Island, Korea.

LI, X., and STERN, R. M. (2004). “Feature Generation Based on Maximum Normalized Acoustic Likelihood for Improved Speech Recognition,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004, Montreal.

RAJ, B., SINGH, R., and STERN, R. M. (2004). “On Tracking Noise with Linear Dynamical System Models,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004, Montreal.

SELTZER, M. L., and STERN, R. M. (2003). “Parameter Sharing in Subband Likelihood-Maximizing Beamforming for Speech Recognition using Microphone Arrays,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004, Montreal.

LI, X., and STERN, R. M. (2003). “Feature Generation Based on Maximum Classification Probability for Improved Speech Recognition," *Proc. Eurospeech-2003*, September, 2003, Geneva, Switzerland.

NEDEL, J. P., and STERN, R. M. (2003). “Duration Normalization and Hypothesis Combination for Improved Spontaneous Speech Recognition,” *Proc. Eurospeech-2003*, September, 2003, Geneva, Switzerland.

OBUCHI, Y., and STERN, R. M. (2003). “Normalization of Time-Derivative Parameters using Histogram Equalization,” *Proc. Eurospeech-2003*, September, 2003, Geneva, Switzerland.

LI, X., and STERN, R. M. (2003). “Training of Stream Weights for the Decoding of Speech using Parallel Feature Streams,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2003, Hong Kong.

SELTZER, M. L., and STERN, R. M. (2003). “Subband Parameter Optimization of Microphone Arrays for Speech Recognition in Reverberant Environments,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2003, Hong Kong.

LI, X., SINGH, R., and STERN, R. M. (2002). "Lattice Combination for Improved Speech Recognition," *Proc. of the International Conference of Spoken Language Processing*, September, 2002, Denver, Colorado.

SELTZER, M. L., RAJ, B., and STERN, R. M. (2002). “Speech Recognizer-Based Microphone Array Processing for Robust Hands-Free Speech Recognition,” Proc. IEEE Conf. on Acoustics, Speech, and Sig. Proc., May, 2002, Orlando, Florida.

RAJ, B., SELTZER, M. L., and STERN, R. M. (2001). “Robust Speech Recognition: The Case for Restoring Missing Features,” Proc. of the Workshop on Consistent and Reliable Auditory Cues, September, 2001, Aalborg, Denmark.

SINGH, R., SELTZER, M. L., RAJ, B., and STERN, R. M. (2001). “Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination,” *Proc. IEEE Conf. on Acoustics, Speech, and Sig. Proc.*, May, 2001, Salt Lake City, Utah.

NEDEL, J. N., and STERN, R. M. (2001). “Duration Normalization for Improved Recognition of Spontaneous and Read Speech via Missing Feature Methods,” *Proc. IEEE Conf. on Acoustics, Speech, and Sig. Proc.*, May, 2001, Salt Lake City, Utah.

DOH, S.-H., and STERN, R. M. (2000). “Using Class Weighting in Inter-Class MLLR,” *Proc. of the International Conference of Spoken Language Processing*, October, 2000, Beijing, China.

HUERTA, J. M., and STERN, R. M. (2000). "Instantaneous Distortion-Based Weighted Acoustic Modeling for Robust Recognition of Coded Speech," *Proc. of the International Conference of Spoken Language Processing*, October, 2000, Beijing, China.

NEDEL, J. P., SINGH, R., and STERN, R. M. (2000a). "Automatic Subword Unit Refinement for Spontanteous Speech Recognition via Phoneword Splitting," *Proc. of the International Conference of Spoken Language Processing*, October, 2000, Beijing, China.

NEDEL, J. P., SINGH, R., and STERN, R. M. (2000b). "Phone Transition Acoustic Modeling: Application to Speaker Independent and Spontaneous Speech Systems," *Proc. of the International Conference of Spoken Language Processing*, October, 2000, Beijing, China.

RAJ, B., SELTZER, M. L., and STERN, R. M. (2000). "Reconstruction of Damaged Spectrographic Features for Robust Speech Recognition," *Proc. of the International Conference of Spoken Language Processing*, October, 2000, Beijing, China.

SELTZER, M. L., RAJ, B., and STERN, R. M. (2000). "Classifier-Based Mask Estimation for Missing Feature Methods of Robust Speech Recognition," *Proc. of the International Conference of Spoken Language Processing*, October, 2000, Beijing, China.

SINGH, R., RAJ, B., and STERN, R. M. (2000). "Structured Redefinition of Sound Units by Merging and Splitting for Improved Speech Recognition," *Proc. of the International Conference of Spoken Language Processing*, October, 2000, Beijing, China.

DOH, S.-J., and STERN, R M. (2000). "Inter-Class MLLR for Speaker Adaptation," *Proc. IEEE Conf. on Acoustics, Speech, and Sig. Proc.*, June, 2000, Istanbul, Turkey.

SINGH, R., RAJ, B., and STERN, R. M. (2000). "Automatic Generation of Phone Sets and Lexical Transcriptions," *Proc. IEEE Conf. on Acoustics, Speech, and Sig. Proc.*, June, 2000, Istanbul, Turkey.

DOH, S.-J., and STERN, R M. (1999). "Weighted Principal Component MLLR For Speaker Adaptation," *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, December, 1999, Keystone, Colorado.

SINGH, R., RAJ, B., and STERN, R. M. (1999). "Domain Adduced State Tying for Cross-domain Acoustic Modelling," *Proc. Eurospeech-99*, September, 1999, Budapest, Hungary.

SINGH, R., RAJ, B., and STERN, R. M. (1999). "Automatic Clustering And Generation Of Contextual Questions For Tied States In Hidden Markov Models," *Proc. IEEE Conf. on Acoustics, Speech, and Sig. Proc.*, March, 1999, Phoenix, Arizona.

HUERTA, J. M., and STERN, R. M. (1999). "Distortion-class weighted acoustic modeling for Robust Speech Recognition under GSM RPE-LTP coding," *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999.

HUERTA, J. M., AND STERN, R. M. (1998). "Speech Recognition from GSM CODEC Parameters," *Proc. of the International Conference of Spoken Language Processing*, Sydney Australia, December, 1998.

RAJ, B., SINGH, R., and STERN, R. M. (1998). "Inference of Missing Spectrographic Features for Robust Speech Recognition," *Proc. of the International Conference of Spoken Language Processing*, Sydney Australia, December, 1998.

RAJ, B., GOUVEA, E., and STERN, R. M. (1997). "Cepstral Compensation using Statistical Linearization," *Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-au-Mousson, France, April, 1997.

GOUVEA, E. B., and STERN, R. M. (1997). "Speaker Normalization through Formant-Based Warping of the Frequency Scale", *Proc. Eurospeech-97*, September, 1997, Rhodes, Greece.

HUERTA, J. M., and STERN, R. M. (1997). "Compensation for Environmental and Speaker Variability by Normalization of Pole Locations," *Proc. Eurospeech-97*, September, 1997, Rhodes, Greece.

RAJ, B, PARIKH, V., and STERN, R. M. (1997). "The Effects of Background Music on Speech Recognition Accuracy," *Proc. IEEE Conf. on Acoustics, Speech, and Sig. Proc.*, April, 1997, Munich, Germany.

RAJ, B., GOUVEA, E., MORENO, P. J., and STERN, R. M. (1996). "Cepstral Compensation by Polynomial approximation for Environment-Independent Speech Recognition," *Proc. of the International Conference on Spoken Language Processing,* Philadelphia, Pennsylvania, October, 1996.

MORENO, P. J., RAJ, B., and STERN, R. M. (1996). "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, 1996.

MORENO, P. J., RAJ, B., and STERN, R. M. (1995). "A Unified Approach to Robust Speech Recognition," *Proc. of Eurospeech-95*, Madrid, Spain, September, 1995.

MORENO, P. J., RAJ, B., GOUVEA, E., and STERN, R. M. (1995). "Multivariate- Gaussian-Based Cepstral Normalization for Robust Speech Recognition," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Detroit, Michigan, 1995.

SIEGLER, M. A., and STERN, R. M. (1995). "On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Detroit, Michigan, May, 1995.

STERN, R. M., LIU, F.-H., MORENO, P. J., and ACERO, A. (1994). "Signal Processing for Robust Speech Recognition," *Proc. of the International Conference on Spoken Language Processing*, Yokohama, Japan, September, 1994.

HANAI, N., and STERN, R. M. (1994). "Robust Speech Recognition in the Automobile," *Proc. of the International Conference on Spoken Language Processing*, Yokohama, Japan, September, 1994.

OHSHIMA, Y., and STERN, R. M. (1994). "Environmental Robustness in Automatic Speech Recognition Using Physiologically-Motivated Signal Processing," *Proc. of the International Conference on Spoken Language Processing*, Yokohama, Japan, September, 1994.

LIU, F.-H., STERN, R. M., ACERO, A., and MORENO, P. J. (1994). "Environment Normalization for Robust Speech Recognition using Direct Cepstral Comparison," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia, pp. II-61 - II-64.

MORENO, P. J., and STERN, R. M. (1994). "Sources of Degradation of Speech Recognition in the Telephone Network," *Proc. of the IEEE International Conference on Acoustics, Speech, and*

Signal Processing, Adelaide, Australia, pp. I-109 - I-112.

LIU, F.-H., MORENO, P. J., STERN, R. M., and ACERO, A. (1994). "Signal Processing For Robust Speech Recognition," *Proceedings of the Seventh ARPA Workshop on Human Language Technology*, Princeton, New Jersey, Morgan Kaufmann, C. J. Weinstein, Ed.

LIU, F.-H., MORENO, P. J., STERN, R. M., and ACERO, A. (1994). "Signal Processing For Robust Speech Recognition," *Proceedings of the ARPA Workshop on Spoken Language Technology*, Princeton, New Jersey, R. M. Stern, Ed.

LIU, F.-H., STERN, R. M., HUANG, X., and ACERO, A. (1993). "Efficient Cepstral Normalization For Robust Speech Recognition," *Proceedings of the Sixth ARPA Workshop on Human Language Technology*, Princeton, New Jersey, Morgan Kaufmann, M. Bates, Ed., pp. 69-74.

SULLIVAN, T. M., and STERN, R. M. (1993). "Multi-Microphone Correlation- Based Processing for Robust Speech Recognition," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, Minnesota, **2**: 91-94.

STERN, R. M., LIU, F.-H., OHSHIMA, Y., SULLIVAN, T. M., and ACERO, A. (1992a). "Multiple Approaches to Robust Speech Recognition", *Proc. of the Fifth DARPA Speech and Natural Language Workshop*, Harriman, New York, February, 1992.

STERN, R. M., LIU, F.-H., OHSHIMA, Y., SULLIVAN, T. M., and ACERO, A. (1992b). "Multiple Approaches to Robust Speech Recognition," *Proc. of the Second International Conference on Spoken Language Processing*, Banff, Alberta, Canada, pp. 695-698, October, 1992.

LIU, F.-H., ACERO, A., and STERN, R. M. (1992). "Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, California, pp. 865-868.

WARD, W., ISSAR, S., HUANG, X., HON, H.-W., HWANG, M.-Y., YOUNG, S., MATESSA, M., LIU, F.-H., and STERN, R. (1992). "Speech Understanding in Open Tasks", *Proc. of the Fifth DARPA Speech and Natural Language Workshop, Harriman*, New York, February, 1992.

STERN, R. M., XU, X., and TAO, S. (1991). "A Coincidence-Based Model that Describes "Straightness" Weighting in Binaural Perception," *Abstracts of the Fourteenth Midwinter Research Meeting of the Association for Research in Otolaryngology*, St. Petersburg Beach, Florida, p. 33(A).

STERN, R. M., and TRAHIOTIS, C. (1991). "The Role of Consistency of Interaural Timing over Frequency in Binaural Lateralization," *Proc. of the Ninth International Symposium on Auditory Physiology and Perception*, Carcans, France.

ACERO, A. and STERN, R. M. (1991). "Robust Speech Recognition by Normalization of the Acoustic Space," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Ontario, pp. 893-896.

ROZZI, W. A. and STERN, R. M. (1991). "Fast Estimation of Mean Vectors using Adaptive Filtering," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Ontario, pp. 865-868.

ACERO, A. and STERN, R. M. (1990a). "Environmental Robustness in Automatic Speech Recognition," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, New Mexico, pp. 849-852.

ACERO, A. and STERN, R. M. (1990b). "Toward Microphone-Independent Spoken Language Systems," *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, R. M. Stern, Ed., Morgan Kaufmann Publishers, Inc., San Mateo, CA,

ACERO, A. and STERN, R. M. (1990c). "Acoustical Pre-Processing for Robust Spoken Language Systems," *Proc. First International Conference on Spoken Language Processing*, pp. 1121-1124, Kobe, Japan, November, 1990.

STERN, R. M., ZEPPENFELD, T., and SHEAR, G. D. (1990). "Lateralization of Rectangularly-Modulated Noise: An Explanation for "Illusory" Reversals," *Abstracts of the Thirteenth Midwinter Research Meeting of the Association for Research in Otolaryngology*, St. Petersburg Beach, Florida, pp. 163-164(A).

STERN, R. M. and ACERO, A. (1989). "Acoustical Pre-Processing for Robust Speech Recognition," presented at the October, 1989, DARPA Workshop on Speech and Natural Language.

WARD, W. H., HAUPTMANN, A. G., STERN, R. M., and CHANAK, T. (1988). "Parsing Spoken Phrases Despite Missing Words," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 275-278.

COAST, D. A., STERN, R. M., CANO, G. C., and BRILLER, S. A. (1987). "Cardiac Arrhythmia Analysis Using Hidden Markov Models," presented at the IEEE Engineering in Medicine Society Conference, November, 1987.

STERN, R. M. (1988). "Nuevos Enfoques en Reconocimiento Automatico de Habla (New Directions in Automatic Speech Recognition)", invited lecture at the XIII Simposium Internacional de Sistemas Computacionales, Monterrey, Mexico.

STERN, R. M., WARD, W. H., HAUPTMANN, A. G., and LEON, J. (1987). "Sentence Parsing with Weak Grammatical Constraints," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 380-383.

LASRY, M.J., AND STERN, R.M. (1984). "Unsupervised Adaptation to New Speakers in Feature-Based Letter Recognition," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 17.6.1-17.6.4.

STERN, R. M., and BACHORSKI, S. J. (1983). "Dynamic Cues in Binaural Perception," in *Hearing-Physiological Bases and Psychophysics*, R. Klinke and R. Hartmann, Eds., Springer Verlag Press, Heidelberg.

STERN, R. M., and LASRY, M. J. (1983). "Dynamic Speaker Adaptation for Isolated Letter Recognition Using MAP Estimation," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 734-737.

COLE, R. A., STERN, R. M., PHILLIPS, M. S., BRILL, S. M., PILANT, A. P., and SPECKER, P. (1983). "Feature-based Speaker-Independent Recognition of Isolated English Letters," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 731-734.

STERN, R. M., Jr. and RUBINOV, E. M. (1980). "Subjective Laterality of Noise-Masked Binaural Targets," in *Psychological, Physiological, and Behavioural Studies of Hearing*, G. v.d. Brink and F. A. Bilsen, Eds., Delft University Press, Delft.

Non-Reviewed Submitted Conference Papers and Other

STERN, R. M. (2017). “Predicting binaural lateralization and discrimination using the position-variable model,” talk at the 173rd meeting of the Acoustical Society of America, Boston, Massachusetts, June, 2018.

DIETZ, M., MARQUARDT, T., MAJDAK, P., STERN, R. M., HARTMANN, W. M., GOODMAN, D. F., and EWERT, S. D. (2017). “An initiative for testability and comparability of binaural models, talk at the 173rd meeting of the Acoustical Society of America, Boston, Massachusetts, June, 2017.

PARK, H.-M., MACIEJEWSKI, M., KIM, C., and STERN, R. M. (2015). “Robust automatic speech recognition in reverberation: onset enhancement versus binaural source separation,” J. Acoust. Soc. Amer., **137**:2303 (A).

ROMIGH, G. D., STERN, R. M., BRUNGART, D. S>, and SIMPSON, B. D. (2015). “A Bayesian framework for the estimation of head-related transfer functions,” J. Acoust. Soc. Amer., **137**:2323 (A).

STERN, R. M. (2003). “Signal Processing for Robust Recognition,” Speech Information Technology Research Center Seminar Series, Seoul, Korea, May, 2003.

NEDEL, J. P., and STERN, R. M. (2002). “Duration Normalization for Improved Automatic Speech Recognition,” *J. Acoust. Soc. Amer.*, **112**: 2321 (A).

HUERTA, J. M., CHEN, S. J., and STERN, R. M. (1999). “The 1998 Carnegie Mellon University Sphinx-3 Spanish Broadcast News Transcription System,” *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, March, 1999, Herndon, Virginia.

HUERTA, J. M., THAYER, E., RAVISHANKAR, M., and STERN, R. M. (1998). “The Development of the 1997 CMU Spanish Broadcast News Transcription System,” *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, February, 1998, Landsdowne, Virginia.

SEYMORE, K., CHEN, S., DOH, S.-J., ESKENAZI, M., GOUVEA, E., RAJ, B., RAVISHANKAR, M. ROSENFELD, R., SIEGLER, M. A., STERN, R. M., AND THAYER, E. (1998). “The 1997 CMU Sphinx-3 English Broadcast News Transcription System,” *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, February, 1998, Landsdowne, Virginia.

SIEGLER, M. A., JAIN, U., RAJ, B., and STERN, R. M. (1997). “Automatic Segmentation, Classification, and Clustering of Broadcast News Audio,” *Proc. DARPA Speech Recognition Workshop*, February, 1997, Chantilly, Virginia.

PARIKH, V. N., RAJ, B., and STERN, R. M. (1997). “Speaker Adaptation and Environmental Compensation for the 1996 Broadcast News Task,” *Proc. DARPA Speech Recognition Workshop*, February, 1997, Chantilly, Virginia.

PLACEWAY, P., CHEN, S., ESKENAZI, M., JAIN, U., PARIKH, V., RAJ, B., RAVISHANKAR, M., ROSENFELD, R., SEYMORE, K., SIEGLER, M., STERN, R., and THAYER, E. (1997). “The 1996 Hub-4 SPHINX-3 System,” *Proc. DARPA Speech Recognition Workshop*, February, 1997, Chantilly, Virginia.

STERN, R. M. (1997). “Specification of the 1996 Hub 4 Broadcast News Evaluation,” *Proc. DARPA Speech Recognition Workshop*, February, 1997, Chantilly, Virginia.

GOUVEA, E. B., MORENO, P. J., RAJ, B., SULLIVAN, T. M., and STERN, R. M. (1996). "Adaptation and Compensation: Approaches To Microphone And Speaker Independence In Automatic Speech Recognition," *Proceedings of the ARPA Workshop on Speech Recognition Technology*, Harriman, NY, Morgan Kaufmann, D. Pallett, Ed.

JAIN, U., SIEGLER, M. A., DOH, S.-J., GOUVEA, E., MORENO, P. J., RAJ, B. and STERN, R. M. (1996). "Recognition Of Continuous Broadcast News With Multiple Unknown Speakers And Environments," *Proceedings of the ARPA Workshop on Speech Recognition Technology*, Harriman, NY, Morgan Kaufmann, D. Pallett, Ed.

STERN, R. M. (1996). "Specification of the 1995 ARPA Hub 3 Evaluation: Unlimited Vocabulary NAB News Baseline," *Proceedings of the ARPA Workshop on Speech Recognition Technology*, Harriman, NY, Morgan Kaufmann, D. Pallett, Ed.

MORENO, P. J., SIEGLER, M. A., JAIN, U., and STERN, R. M. (1995). "Continuous Recognition of Large- Vocabulary Telephone-Quality Speech," *Proceedings of the ARPA Workshop on Spoken Language Technology*, Austin, TX, Morgan Kaufmann, J. Cohen, Ed.

MORENO, P. J., JAIN, U., RAJ, B., and STERN, R. M. (1995). "Approaches to Microphone Independence in Automatic Speech Recognition," *Proceedings of the ARPA Workshop on Spoken Language Technology*, Austin, TX, Morgan Kaufmann, J. Cohen, Ed.

LEE, W., and STERN, R. M. (1994). "Consistency Over Frequency in High- Frequency Binaural Lateralization," *J. Acoust. Soc. Amer.*, **95**: 2896 (A).

STERN, R. M., and SULLIVAN, T. M. (1993). "Microphone-Array Algorithms for Robust Speech Recognition," talk at the Speech Research Symposium-XIII, Baltimore, Maryland, June, 1993.

TAO, S. H., and STERN, R. M. (1992). "Additive versus Multiplicative Combination of Differences of Interaural Time and Intensity," *J. Acoust. Soc. Amer.*, 91: 2414(A).

STERN, R. M., LIU, F.-H., OHSHIMA, Y., SULLIVAN, T. M., and ACERO, A. (1992c). "Alternative Approaches to Acoustical Pre-Processing for Robust Speech Recognition," presented at the Speech Research Symposium-XII, June, 1992.

ACERO, A., and STERN R. M. (1992). "Cepstral Normalization for Robust Speech Recognition," *Proc. of the ESCA Workshop on Speech Processing in Adverse Conditions*, Cannes-Mandelieu, France, pp 89-92.

STERN, R. M., and ACERO, A. (1990a). "Acoustic Pre-Processing for Robust Spoken Language Systems," invited presentation at the Speech Research Symposium- X, Baltimore, MD, October 16, 1990.

STERN, R. M., and ACERO, A. (1990b). "Signal Processing for Robust Spoken Language Systems," invited tutorial lecture at the Satellite Workshop on Statistical Approaches to Spoken Language Processing, Tokyo, Japan, November 17, 1990.

STERN, R. M., and RUDNICKY, A. I. (1990). "Spoken-Language Workstations in the Office Environment," invited lecture at the SpeechTech90 Conference, New York City.

STERN, R. M., SHEAR, G. D., and ZEPPENFELD, T. (1988). "Lateralization Predictions for High-Frequency Binaural Stimuli," *J. Acoust.Soc. Amer*. **84**: S80 (A).

SHEAR, G. D., and STERN, R. M. (1987). "Extending the Position-Variable Model: Dependence

of Lateralization on Various Spectral Parameters," *J. Acoust. Soc. Amer*. **81**: S27 (A).

STERN, R. M., and ZEIBERG, A. S. (1986). "A Weighted Image Model for Binaural Lateralization," *J. Acoust. Soc. Amer.* **80**: S107 (A).

BEECHER, L., and STERN, R. M. (1986). "Perception of Modulations in Pitch and Lateralization," *J. Acoust. Soc. Amer*. **79**: S22 (A).

STERN, R.M., ELSNER, A.E., and SCHIANO, J.L. (1984). "Interaural Time Discrimination in Tonal Maskers," *J. Acoust. Soc. Amer*. **76**: S91(A).

LASRY, M. J., and STERN, R. M. (1984). "Unsupervised Speaker Adaptation in Feature-Based Isolated Letter Recognition," *J. Acoust. Soc. Amer*. **74**: S16(A).

BACHORSKI, S. J., and STERN, R. M. (1983). "Dynamic Cues in Binaural Perception," *J. Acoust. Soc. Amer*. **73**: S42(A).

STERN, R. M. and LASRY, M. J. (1982). "Tuning to the Speaker: Dynamic Adaptation of Statistical Parameters in Isolated Letter Recognition," *J. Acoust. Soc. Amer*. 72 S31(A).

BRILL, S. M., PHILLIPS, M. S., LASRY, M. J., and STERN, R. M. (1982). "Decisions about Features," *J. Acoust. Soc. Amer*. 72 S32(A).

KAISER, D. L., and STERN, R. M. (1981). "Interaural Time Discrimination in Tonal Maskers," *J. Acoust. Soc. Amer*. 70, S88 (A).

SLOCUM, J. E., and STERN, R. M., Jr., (1980). "Interaural Time and Amplitude Discrimination in Noise," *J. Acoust. Soc. Amer*. 68, S60(A).

STERN, R. M., Jr. and WAIBEL, A. H. (1980). "Audibility of Phase Changes in Vowel Sounds and Complex Tones," *J. Acoust. Soc. Amer*. 68, S50(A).

DUMOND, G. J. and STERN, R. M., Jr. (1979). "A Forced-Choice Paradigm for Pulsation-Threshold Measurements," *J. Acoust. Soc. Amer*. 65, S58(A).

RUBINOV, E. M. and STERN, R. M., Jr. (1979). "Effects of Binaural Maskers on the Subjective Laterality of Diotic Targets," *J. Acoust. Soc. Amer.*, 65, S121(A).

STERN, R. M., Jr. (1979). "On the Use of Multiple Perceptual Images in Binaural Discrimination Experiments," *J. Acoust. Soc. Amer*. 65, S122(A).

RUOTOLO, B. R., STERN, R. M., Jr., and COLBURN, H. S. (1977). "Discrimination of Symmetric, Time-Intensity Traded Stimuli," *J. Acoust. Soc. Amer.*, 61, S60(A).

STERN, R. M., Jr. (1977). "Lateralization of the MLD: Detection-Threshold Performance of an Auditory-Nerve-Based Model for Lateral Position," *J. Acoust. Soc. Amer.*, 61, S60(A).

COLBURN, H. S., DOMNITZ, R. H., STERN, R. M., Jr., and DURLACH, N. I. (1976). "Current Problems in Binaural Hearing Research," *J. Acoust. Soc. Amer.*, 59, S16(A).

STERN, R. M., Jr. (1976a). "Lateral Position, Interaural Discrimination, and Binaural Detection: Model Based on Auditory-Nerve Activity," *J. Acoust. Soc. Amer.*, 59, S23(A).

STERN, R. M., Jr. (1972). "Perception of Simultaneously-Presented Musical Timbres," *Quart.*

Prog. Rep. No. 106, Res. Lab. of Electronics, MIT, Cambridge, MA.

Ph.D. THESES SUPERVISED

D. A. Coast, *Cardiac Arrhythmia Analysis Using Hidden Markov Models*, September, 1988.

A. Acero, *Acoustical and Environmental Robustness for Automatic Speech Recognition*, September, 1990.

W. A. Rozzi, *Speaker Adaptation in Automatic Speech Recognition via Estimation of Correlated Mean Vectors*, May, 1991.

Y. Ohshima, *Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing*, December, 1993.

F.-H. Liu, *Environmental Adaptation for Robust Speech Recognition*, June, 1994.

P. J. Moreno, *Speech Recognition in Noisy Environments*, May, 1996.

T. M. Sullivan, *Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition*, June, 1996.

E. Gouvea, *Acoustic-Feature-Based Frequency Warping for Speaker Normalization,* February, 1999.

M. Siegler, *Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance*, December, 1999.

B. Raj, *Reconstruction of Incomplete Spectrograms for Robust Speech Recognition,* April, 2000.

J. Huerta, *Robust Speech Recognition in GSM Codec Environments*, April, 2000.

S.-J. Doh, *Enhancements to Transformation-Based Speaker Adaptation: Principal Component and Inter-Class Maximum Likelihood Linear Regression*, July, 2000.

M. Seltzer, *Microphone Arrays for Robust Speech Recognition,* July, 2003.

J, Nedel, *Duration Normalization for Robust Recognition of Spontaneous Speech via Missing Feature Methods,* April, 2004.

X. Li, *Combination and Generation of Parallel Feature Streams for Improved Speech Recognition*, February, 2005.

Z. Al Bawab, *An Analysis-by-Synthesis Approach to Vocal Tract Modeling for Robust Speech Recognition,* September, 2009.

Y.-H. Chiu, *Learning-Based Auditory Encoding for Robust Speech Recognition,* April, 2010.

L. Gu, *Single-Channel Speech Separation Based on Instantaneous Frequency,* May, 2010.

C. Kim, *Signal Processing for Robust Speech Recognition Motivated by Auditory Processing*, September, 2010.

K. Kumar, *A Spectro-Temporal Framework for Compensation of Reverberation for Speech Recognition,* January, 2011.

G. Romigh, *Individualized Head-Related Transfer Functions: Efficient Modeling and Estimation from Small Sets of Spatial Samples*, December, 2012.

A. Moghimi, *Array-Based Spectro-Temporal Masking for Automatic Speech Recognition,* April, 2014.

M. J. Harvilla, *Compensentation for Nonlinear Distortion in Noise for Robust Speech Recognition,* October, 2014.

A. Menon, *Robust Recognition Of Binaural Speech Signals Using Techniques Based On Human Auditory Processing*, February, 2019.

CURRENT Ph.D. STUDENTS

Y. Xia, entered August 2016.

T. Vuong, entered August 2017.

P. Conrey, entered August 2019.

M. Lindsay, entered August 2019.

M.S. THESES and PROJECTS SUPERVISED

K. G. Murti, *Sonometric Evaluation of Eustacian-Tube Function*, May, 1979.

G. J. DuMond, *A Forced-Choice Paradigm for Pulsation-Threshold Measurement of Monaural and Binaural Phenomena*, August, 1979.

E. M. Rubinov, *Auditory Lateralization in Noise*, September, 1979.

J. E. Slocum, *Discrimination of Interaural Time and Intensity in Noise*, November, 1980.

D. L. Kaiser, *Interaural Time Discrimination in Tonal Maskers*, June, 1982.

A. H. Ferguson, *Effects of Frequency Modulation on Complex Pitch Perception*, December, 1982.

M. J. Lasry, *Dynamic Adaptation of Statistical Parameters in a Feature-Based Isolated Letter Recognition System*, December, 1982.

Y. L. Chu, *Zero Crossings and Ear Modelling in Speech*, February, 1983.

S. J. Bachorski, *Dynamic Cues in Binaural Perception*, December, 1983.

W. J. Lue, *Identification of Transient Features in Speech*, January, 1984.

L. E. Beecher, *Perception of Modulations in Pitch, Pitch Strength, and Subjective Lateral Position*, December, 1986.

S. J. Bartschat, *Bayesian Combination of Knowledge Sources in Automatic Speech Recognition*, August, 1987.

M. P. Vea, *Multisensor Signal Enhancement for Speech Recognition*, September, 1987.

G. D. Shear, *Modelling the Dependence of Auditory Lateralization on Frequency and Bandwidth*, September, 1987.

B. Chigier, *Classification of Stop Consonants in Natural Continuous Speech*, May, 1988.

P. H. Dietz, *Simulation of Trumpet Tones Via Physical Modeling*, December, 1988.

S. Palm, *Enhancement of Reverberated Speech using Models of the Human Binaural System*, April, 1989.

A. Liss, *Lateralization of Complex Binaural Sounds*, July, 1990.

S. H. Tao, *Additive versus Multiplicative Combination of Differences of Interaural Time and Intensity*, May, 1992.

P. J. Moreno, *Speech Recognition in Telephone Environments*, January, 1993.

J. Nehme, *Robust Recognition of Speech Input to Computer Workstations*, March, 1993.

N. Hanai, *Speech Recognition in the Automobile*, May, 1993.

W. Lee, *Lateralization of High-Frequency Bandpass Noise*, May, 1994.

M. Siegler, *Effects of Speech Rate on Speech Recognition Accuracy*, December, 1995.

U. Jain, *Speech Recognition over the Telephone*, May, 1995.

J. Nedel, *Integration of Speech and Video. Applications for Lip Synch: Lip Movement Synthesis and Time Warping,* May, 1999.

A. Ripepi, *Lateralization and Identification of Simultaneously-Presented Whispered Vowels and Speech Sounds,* May, 1999.

M. Seltzer, *Automatic Detection of Corrupted Speech Features for Robust Speech Recognition*, May, 2000.

M. Balakrishnan, *Extracting Additional Information from Gaussian Mixture Probabilities for Unsupervised Text-Independent Speaker recognition,* May 2005.

D. Zurale, *TaanSyn - A Synthesizer to add critical Indian Classical Singing elements to Western Vocals,* May 2017.

J. Michelson, *Unsupervised Guitar String Classification for Tablature Transcription*, August 2017.

C. Liang, *A Comparative Evaluation of Statistical Models for Score Following,* October 2017.

M. Yu. *How Important are the Headphone and Ear Canal Frequency Responses in Music Perception?* October, 2018.

Y. Zhang. *Design of Matching Criteria for Audio-Based Polyphonic Score-following Systems Using Harmonic Product Spectra,* December 2018.

COURSES TAUGHT AT CARNEGIE MELLON UNIVERSITY

18-101	Linear Circuit Analysis (Sophomore)
18-290	Signals and Systems (Sophomore and Junior)
18-396	Signals and Systems (Junior)
18-301	Fundamentals of Control (Senior)
18-308	Introduction to Digital Communication (Senior)
18-491	Introduction to Signal Processing (Junior/Senior)
18-493	Electroacoustics (Senior)
18-550	Fundamentals of Communication (Senior)
18-551	Communication and Signal Processing Project Laboratory (Senior)
18-340	Senior Seminar on Speech Processing
18-762	Signal Transmission and Detection (Graduate)
18-791	Digital Signal Processing (Graduate)
18-792	Advanced Digital Signal Processing (Graduate)
18-799	Digital Signal Processing for Robust Speech Recognition (Graduate)
18/42-795	Sensory Processes: Perception and Psychophysics (Graduate/ Senior)
42-726	Hearing: Physiology and Perception (Graduate Seminar)

(18-308, 18-493, 18-791, 18-792, 18-795, and 42-726 were all new courses when first taught)

SERVICE ON DEPARTMENT AND UNIVERSITY COMMITTEES

Major contributions and initiatives:

Developed and obtained university approval for the CMU Programs in Music and Technology. Service on the CMU Music and Technology Executive Committee, 2009 - present.

As Associate Director of the CMU Information Networking Institute managed the MSIN master's program that spanned the engineering, computer science, management, and public policy colleges, and had primary responsibility for admissions and recruiting, 1994 - 2003.

Chair, Faculty Senate Intellectual Property Policy Committee, 1983 - 1986. Drafted and negor-

tiated the approval of the intellectual property policy now in use by CMU. Chairing committee to review and recommend modifications to the Policy, 2017 - present.

Faculty Senate International Affairs Forum, 1989 - 1991. Initiated cooperative educational programs with Latin American universities.

Chair and Senate Delegate, Faculty Senate Student Affairs Council, 1981 - 1984. Drafted CMU alcohol policy and set faculty policy on a range of other student-related issues.

Other activities (partial listing):

Faculty Senate Executive Committee

University Trustees Student Affairs Council

University Education Council

University Research Council

University Review Committee for Non-Tenured Appointments

Senate Delegate, Faculty Senate Student Affairs Council

Senate Delegate, Faculty Senate Educational Affairs Council

International Committee of the Dean's Council

Carnegie Institute of Technology Faculty Chair

MISCELLANEOUS SKILLS AND INTERESTS

Solo and chamber music performance on harpsichord with ensembles playing historical instruments, and with members of the Pittsburgh Symphony Orchestra. Participation in mountain biking, running, backpacking, racquet sports.

Extensive travel experience throughout Europe, Latin America, and Asia. Fluent in Spanish; some French, Italian, and German.